Sumário

What is “More than Moore”?

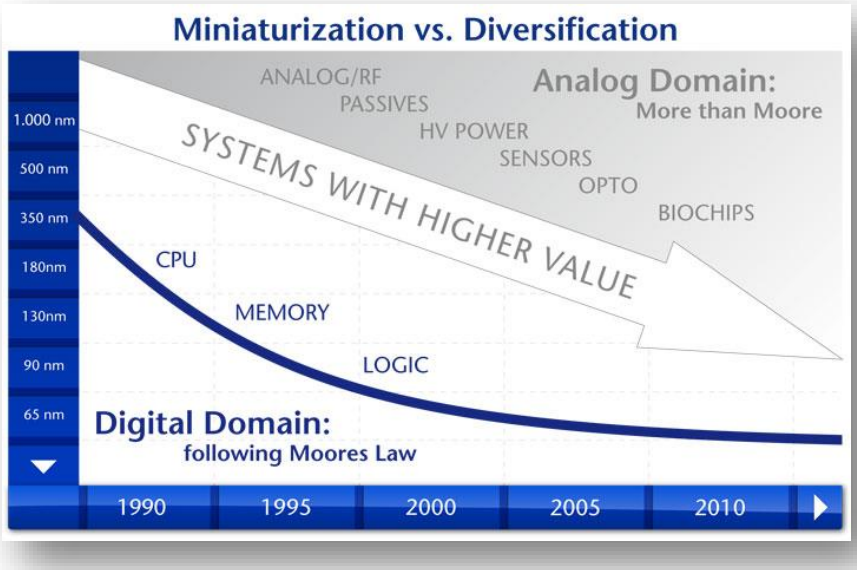
Fonte: https://www.xfab.com/about-x-fab/more-than-moore/what-is-more-than-moore/

O que é "Mais que Moore"?

A Lei de Moore (o número de transistores em um chip dobra a cada 18 a 24 meses) alimentou a microeletrônica convencional nas últimas décadas, reduzindo os ICs para 45 nm e abaixo e prometendo custos mais baixos para os fabricantes de chips. Essa escala extrema funciona bem para memórias e microprocessadores no mundo digital, mas não para fazer interface com o mundo físico real, que é analógico.

Muitas aplicações, como dispositivos de radiofrequência (RF), subsistemas de gerenciamento de energia, componentes passivos, biochips, sensores, atuadores, sistemas microeletromecânicos (MEMS[1]), desempenham um papel igualmente importante nos produtos semicondutores atuais. A integração de funções analógicas em tecnologias especializadas baseadas em CMOS permite soluções de sistema com custo otimizado e valor agregado. Essas tecnologias diversificadas são conhecidas como "Mais que Moore".

O ecossistema de design da X-FAB vai além das soluções para a lógica e o dimensionamento da memória para fornecer valor "Mais que Moore" para os clientes. Em vez de seguir a Lei de Moore, o X-FAB integra recursos de tecnologia que interagem com o mundo analógico e fornece um ecossistema de design abrangente. Inclui serviços e ferramentas para o desenvolvimento de produtos diversificados de energia / HV, MEMS, opto e analógico; um serviço de linha direta técnica 24 horas; um portfólio de bibliotecas extensas tecnicamente maduras e IP; um amplo espectro de dispositivos primitivos; e opções flexíveis de prototipagem. Todos são apoiados pelos 20 anos de experiência sólida da X-FAB em fundição analógica / sinal misto. O resultado - um ecossistema de design com amplas soluções "More than Moore" - liga idealmente os mundos digital e analógico e fornece valor ideal para os clientes.



More than Moore Devices: The Wind of Change

Fonte: https://blog.semi.org/business-markets/more-than-moore-devices-the-wind-of-change

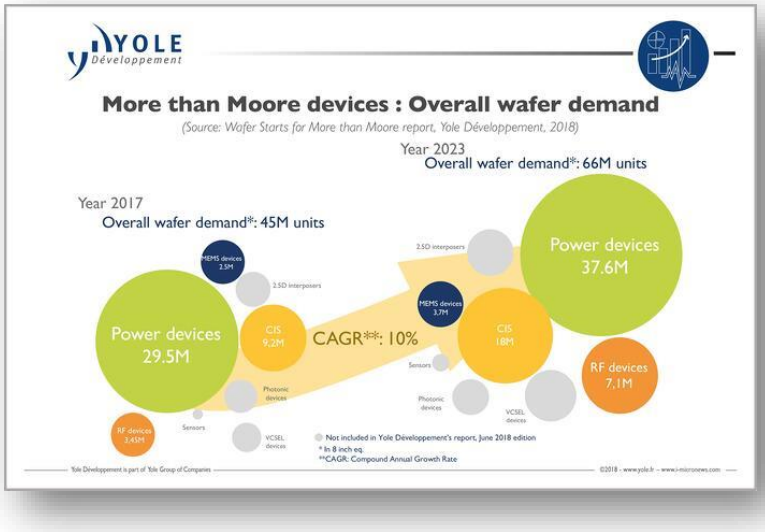
Mais do que Moore Devices: O vento da mudança

Impulsionada pela adoção de cada vez mais componentes eletrônicos em produtos finais, a indústria de semicondutores está enfrentando uma nova era na qual o dimensionamento de dispositivos e a redução de custos não continuarão mais no caminho seguido nas últimas décadas. Os nós avançados não trazem mais o benefício de custo desejado, e os investimentos em P&D em novas soluções e dispositivos de litografia com menos de 10nm estão aumentando substancialmente. Para atender às demandas do mercado, a indústria está procurando soluções tecnológicas para preencher a lacuna e melhorar o custo / desempenho, além de adicionar mais funcionalidade por meio da integração.

Os dispositivos mais do que Moore (incluindo MEMS e sensores, sensores de imagem CMOS, eletrônicos de potência e dispositivos de RF) representam essa nova diversificação funcional de tecnologias, combinando desempenho, integração e custo não limitados ao dimensionamento de CMOS, e sua importância se tornará cada vez maior e mais preponderante.

Em 2017, a demanda de wafer para dispositivos mais que Moore atingiu quase 45 milhões de eq. de 8 polegadas wafers. Espera-se que esse número atinja mais de 66 milhões de wafers eq. de 8 polegadas até 2023, mostrando um crescimento de quase 10% durante esse período.

Esse aumento é suportado pelas famosas megatendências detalhadas na nova análise, Wafer Starts for More Than Moore Applications, realizada por Yole Développement (Yole). Essa análise é relevante para os seguintes mercados: 5G com infraestrutura sem fio e segmentos móveis, incluindo funcionalidades adicionais, processamento de voz, automotivo inteligente e eletrificação, AR/VR e AI (Artificial Intelligence).



Pela primeira vez, a empresa de consultoria de pesquisa e estratégia de mercado apresenta uma tecnologia dedicada e uma análise de mercado focada na demanda geral de wafer por dispositivos mais que Moore. O objetivo deste relatório é fornecer uma visão geral das

remessas de wafer para dispositivos mais que Moore, do tamanho da wafer ao tipo de substrato do material semicondutor, incluindo silício, vidro, SOI (Silicon on Insulator), SiC (Silicon Carbid), SiGe (Silicon Germanium), GaN (Gallium Nitride), InP (Indium Phosphide), GaAs (Gallium Arsenide), safira e cerâmica e, portanto, identificar oportunidades de negócios no setor mais que Moore.

Por mais de 20 anos, a Yole analisa a evolução do setor, discute com empresas líderes para entender os desafios do mercado e identifica avanços técnicos. O relatório Wafer Start for More Than Moore Applications é o resultado dessa pesquisa de 20 anos. Os analistas da Yole combinam conhecimento técnico e de mercado para descrever o mundo de mais que Moore. Tamanho do mercado (volume e valor), tamanhos e formatos de substratos, cadeia de valor, processos de tecnologia e direcionadores de mercado, oportunidades de negócios e cenário competitivo fazem parte da análise da Yole.

As várias equipes de pesquisa da Yole, abrangendo eletrônica de potência, geração de imagens e sensoriamento, fabricação de RF e semicondutores, colaboram para apresentar uma compreensão aprofundada da atual evolução do mercado, levando em consideração inovações e negócios emergentes. Essa metodologia permite que a Yole cubra as megatendências gerais e ilustre os links entre substrato de wafer, dispositivo, módulo, subsistema, sistema e produto de ponta.

Sob esse ecossistema dinâmico, a implantação de fontes renováveis de energia e acionamentos de motores industriais, bem como a eletrificação da indústria automotiva, são bons exemplos do impacto das megatendências no desenvolvimento da indústria de semicondutores. Eles estão impactando diretamente o mercado de wafer dos dispositivos de energia, resultando em um CAGR[1] esperado de 13% entre 2017 e 2023. Já em 2017, esse mercado representava mais de 60% do mercado geral de wafer para dispositivos mais que Moore e atualmente ainda está dominando a indústria mais que Moore.

5G é uma megatendência que impulsiona a demanda de wafer. A 5G está liderando a evolução mais que Moore, levando qualquer serviço a qualquer usuário, em qualquer lugar. As antenas e as funcionalidades de filtragem são duas das principais inovações dessa evolução.

Sem dúvida, os requisitos rigorosos do 5G estão aumentando a demanda por componentes de RF como filtros de RF, amplificadores de potência (PAs) e amplificadores de baixo ruído (LNAs) para garantir o acesso à rede de rádio de amanhã.

Este ano, a Corning e a Menlo Micro anunciaram um grande acordo para desenvolver uma plataforma de produtos DMS (Digital-Micro-Switch). Ambos os parceiros propõem uma abordagem inovadora baseada na tecnologia de embalagem TGV (Through Glass Via). Segundo os dois parceiros, essa opção técnica permite cobrir a operação de frequências além de 50GHz.

Entre as numerosas megatendências, a mobilidade não está muito atrás do 5G. A demanda por aplicativos móveis avançados que integram cada vez mais funcionalidade está crescendo. Para competir, as empresas estão desenvolvendo combinações inteligentes de dispositivos, como sensores de impressão digital, sensores de luz ambiente, detecção 3D, microfones e dispositivos MEMS inerciais. Como exemplo, desenvolvimentos impressionantes focados em sensores NIR baseados em SOI foram lançados pela SOITEC para aplicativos de

criação de imagens na parte frontal, incluindo avançados sensores de imagem 3D. Essa evolução técnica contribuirá claramente, em um futuro próximo, para um forte crescimento do mercado de wafer para MEMS e sensores.

Além disso, a indústria automotiva, com o desenvolvimento de carros inteligentes, alcançou um novo nível de complexidade, exigindo o desenvolvimento e a integração de novos sensores. Nesse contexto, muitas empresas têm como objetivo ampliar seus recursos no ADAS (Autonomous Driving Assistance Service) e na direção autônoma. Recentemente, a empresa líder On Semiconductor adquiriu a SensL Technologies, líder em produtos de sensoriamento SPAD e LiDAR para automóveis. Essa aquisição é um sinal entre muitos, destacando a evolução dessa indústria histórica, buscando novos conhecimentos e acolhendo novos players, mais conscientes dos hábitos e necessidades dos consumidores.

Os analistas da Yole esperam que os automóveis inteligentes gerem um crescimento consistente da produção de CIS[3] e wafer de sensor nos próximos cinco anos. É alimentado pela crescente integração de módulos sensores de alto valor como RADAR, geração de imagens, LiDAR e muito mais. Embora o setor automotivo seja apoiado principalmente por essas áreas de crescimento, MEMS e sensores históricos, como sensores de pressão MEMS e MEMS inercial, continuarão crescendo a uma taxa razoável, apoiando o mundo automotivo padrão.

The End of Moore's Law

Fonte: https://www.labs.hpe.com/pdf/CISE-19-02-Williams.pdf

O fim da lei de Moore

Richard Stanley Williams | Hewlett Packard Labs

O fim da lei de Moore pode ser a melhor coisa que aconteceu na computação desde o início da lei de Moore. Confrontar o final de uma época deve permitir uma nova era de criatividade, incentivar cientistas da computação a inventar dispositivos, circuitos e arquiteturas biologicamente inspirados implementado usando tecnologias emergentes recentemente.

Todos sabíamos que estava chegando, mas ninguém que eu conhecia adivinhou que levaria tanto tempo. Em 1990, eu disse que seria 2000 por causa da transparência do óxido do portão. Em 2000, eu disse que seria em 2010 por causa do custo proibitivo de uma fab. Em algum momento, os fatores físicos fundamentais do limite do tamanho de um átomo garantiriam uma parada difícil em algum lugar. Agora, o consenso aparece é que o limite final de tamanho utilizável do comprimento da porta de um transistor de efeito de campo será de 5 nm, o que é essencialmente a largura de nove células unitárias de cristal de silício. Dada a trajetória atual do setor, essa geração do CMOS estará em produção na próxima década. Ainda há quem preveja que 5 nm CMOS será muito caro, consumirá muita energia ou não será confiável o suficiente para atingir a onipresença e o impacto das gerações anteriores. Desisti de apostar contra a engenhosidade de engenheiros determinados e criativos: CMOS de 5 nm é inevitável.

Manter o crescimento exponencial em uma métrica tecnológica por mais de 50 anos tem sido pouco milagroso, mas essa conquista teve um custo. O esforço para escalar o CMOS de silício predominantemente tem dominado os investimentos de capital intelectual e financeiro da indústria, governo e academia, investigações famintas em amplos segmentos da ciência da computação e travando em um modelo dominante para computadores, a arquitetura von Neumann. As proezas de redução do processo sobrecarregaram a inovação no nível dos sistemas, forçar o mercado a se adaptar a uma tecnologia fixa em vez de ter tecnologias que atendam às necessidades do mercado.

Comecei a me preparar para o fim da lei de Moore há 20 anos, e o que percebi é que não é simples, a mudança tecnológica nos colocará em uma nova curva de escala exponencial para o próximo meio século.

Todo avanço a partir de agora pode exigir descobertas significativas e invenções em toda a amplitude física, ciências biológicas e da computação, ou essencialmente um novo milagre a cada dois anos. O trabalho brutalmente duro está apenas começando, e de muitas maneiras que faz agora o momento mais emocionante para trabalhar nas áreas de eletrônica e computação desde os dias de Shannon, Turing, Bardeen e von Neumann.

Não podemos apenas?

Discussões sem fim e inúmeras oficinas tentaram abordar a questão "O que vem a seguir?". Não podemos encontrar um substituto para os transistores CMOS que podemos simplesmente inserir na nossa cadeia de suprimentos existente? Certamente deve haver alguma maneira de continuar a extrair mais desempenho e maior eficiência de um circuito integrado? Que tal transistores da nova era que usam materiais diferentes (GaAs, alguém?) Ou processos físicos (torque de rotação?), deixando-nos construir circuitos com velocidade de ordem de grandeza maior com energia operacional correspondentemente menor que o Si? Houve muitas propostas interessantes nesse sentido, mas pelo menos até agora, uma análise detalhada de Dmitri Nikonov e Ian Young da Intel indica que nada proposto até o momento excederá o desempenho do CMOS escalado o suficiente para fazer uma diferença significativa no nível do circuito.

Não podemos encontrar uma maneira melhor de usar as etapas do processo CMOS que já temos? Talvez a resposta seja a integração ou empilhamento 3D para atingir um enorme paralelismo. No entanto, parece-me improvável que a indústria de semicondutores seja capaz de dobrar continuamente o número de camadas em uma pilha a cada dois anos ou mais até 10^6 para atingir densidades efetivas de transistor de 10^{18} / cm^2 (mas veja acima o sucesso de minhas previsões anteriores). Mesmo que pudessem ser construídos, esses "tijolos" exigiriam operação adiabática ou quase reversível para reduzir drasticamente a energia por operação para impedir que os circuitos se vaporizassem quando energizados. Isso é possível, mas o problema é que ele requer velocidades de clock muito lentas que desperdiçam a luta para alcançar a integração 3D em escala. Certamente, haverá mais aplicações de integração multicamada (isto é, 2,5 D) e chips empilhados, mas essa abordagem é improvável para fornecer melhorias exponenciais ao longo de décadas.

Não podemos simplesmente complementar o CMOS com computação mais exótica, como a quântica ou a óptica? Eu vejo esses nichos importantes algum dia, provavelmente em algumas décadas, mas não no mainstream. Outra alternativa é: "Não podemos nos dar bem dentro dos limites do terminal CMOS de 5 nm?" O problema dessa abordagem é que enfrentamos demandas exponencialmente crescentes de computação, mas a energia disponível na rede será fixada por algum tempo. Até o final da década, haverá 8 bilhões de pessoas na Terra, transportando 20 bilhões de dispositivos móveis e interagindo com 100 bilhões de agentes inteligentes, fixos e móveis autônomos. Fornecer comida, água, educação e oportunidades de emprego que se ajustem à dignidade individual de cada pessoa é um desafio econômico e tecnológico sem precedentes. A Internet das Coisas e a análise de big data prometem solucionar alguns dos graves problemas que enfrentamos, mas os dados gerados estão crescendo exponencialmente mais rápido do que nossa capacidade de os transformar em informações. Qualquer resposta linear a um desafio exponencial falhará.

Hora da mudança

Como eu disse, não há soluções fáceis: acabaremos usando todos os truques que pensamos e procuraremos continuamente por mais para dimensionar nossa computação exponencialmente no futuro. Quanto mais longe podemos ir? Em princípio, ainda existem muitas melhorias de ordem de magnitude em termos de custo de energia por operação lógica, de transporte ou de apagamento disponíveis antes que a computação chegue a qualquer limite físico fundamental, se é que existe.

Um conjunto coerente e interdependente de mudanças significativas em máquinas e algoritmos será necessário para manter uma escala exponencial estendida na computação. Eles não podem ser criados isoladamente; não temos o luxo de buscar um desenvolvimento paralelo independente, isolado da infraestrutura existente. Além disso, essas mudanças deverão ser cuidadosamente realizadas, porque mesmo que ocorram mudanças revolucionárias no hardware do computador, elas precisarão parecer mudanças evolutivas para programadores e desenvolvedores de aplicativos. Aqueles que trabalham para construir novos computadores precisam entender que os usuários finais não podem parar tudo o que estão fazendo para reescrever todo o código legado, nem podemos esperar até que uma geração totalmente nova de cientistas da computação seja treinada em um currículo transformado. Assim, novos sistemas devem ser capazes de executar aplicativos existentes e fazê-lo fora da caixa significativamente melhor do que os sistemas atuais para justificar o investimento em novo hardware. O objetivo de uma substituição ou substituição de tecnologia precisa ser que o novo sistema seja capaz de um desempenho dramaticamente melhor do que a tecnologia antiga, à medida que os programadores aprendem a tirar proveito de recursos adicionais, que devem ser o mais fácil possível de usar e ao mesmo tempo o tempo permite novos aplicativos que antes eram impossíveis. Esta não é uma nova visão: o hardware e o software sempre tiveram que co-evoluir ou jogar o salto para que ambos pudessem melhorar juntos.

No entanto, também existem mudanças fundamentais na natureza dos cálculos que as pessoas desejam realizar. Em vez da computação tradicional de alto desempenho, que exige aritmética de precisão estendida para calcular uma resposta para 18 ou mais números

significativos, ou computação corporativa dedicada ao desempenho como sistemas de registro para lógica de negócios transacionais, as cargas de trabalho dominantes de hoje envolvem frequentemente a pesquisa em terabytes ou mais de dados armazenados e streaming de dados para encontrar algo de significativo. A métrica para classificar um computador de alto desempenho está mudando de FLOPS (operações de ponto flutuante por segundo) para TEPS (arestas atravessadas por segundo) ou GUPS (atualizações de giga por segundo). Provavelmente, a única maneira de melhorar significativamente a produtividade e a eficiência computacional no futuro será projetar máquinas otimizadas para uma tarefa específica; após a lei de Moore, há poucas esperanças de que as melhorias nos processadores de uso geral acabem produzindo melhorias em todos os softwares existentes. Em vez de máquinas de uso geral que contêm muitas unidades de processamento idênticas, precisaremos projetar e construir de forma automática e econômica sistemas sob medida que contenham elementos de computação heterogêneos que possam trabalhar juntos em uma interconexão plug-and-play em resposta a uma necessidade do cliente / aplicativo. Assim, prevejo que veremos a tendência oposta no design de chips das últimas décadas. A perseguição obstinada à lei de Moore levou ao aumento da captura e integração de funções em chips de uso geral, porque o aumento no desempenho da próxima geração do CMOS fez com que os investimentos não recorrentes envolvidos no projeto e na fabricação de chips de propósito especial fossem rapidamente perdidos. Por outro lado, com o final do dimensionamento do transistor, as vantagens de eficiência e preço irão para novos tipos de ASICs como aceleradores que são otimizados para uma função de computação específica e custo de fabricação. Para que essa abordagem se torne realidade, deve haver uma interface aberta e padrões da indústria que permitam que as tecnologias heterogêneas se comuniquem com alta largura de banda e baixa latência, como o tecido Gen-Z (http://genzconsortium.org) O excesso de capacidade industrial nas penúltimas etapas ou ainda mais antigas da escala do processo pode ser reutilizado para democratizar a inovação, permitindo que uma ampla gama de concorrentes contribua para o avanço da computação e conduza a uma nova economia de otimização.

Assim, lançar uma nova era na computação exigirá uma plataforma inovadora e flexível. Um exemplo é um programa de pesquisa iniciado no Hewlett Packard Labs chamado The Machine. A seguir, descrevo alguns dos atributos desta plataforma: computação orientada à memória utilizando principalmente memória não volátil em vez de DRAM e discos rígidos, uma malha fotônica que se conecta no nível do chip e um ambiente de processamento heterogêneo no qual o programa é enviado ao dados. Quando essa plataforma base existe, ela pode ser aprimorada continuamente, anexando aceleradores de propósito específico ou ASICs ao tecido existente do tipo Gen-Z, que executa determinadas tarefas com ordens de magnitude de maneira mais rápida e eficiente do que o CMOS padrão, o que fornecerá avanços líquidos significativos para aplicações particulares. Um lugar para encontrar inspiração para esses aceleradores é nos tipos de computação realizados no cérebro. Por fim, especulo que os computadores mais rápidos e com maior eficiência energética podem ser aqueles que não calculam uma resposta para um problema - em vez disso, podemos ensinar aos computadores como adivinhar uma resposta e confirmá-la ou procurar respostas, se já existir quando necessário.

Computação orientada a memória

Há quase uma década, participei de um exercício patrocinado pela DARPA, liderado por Peter Kogge, para determinar o que seria necessário para construir um computador de escala exótica (que em princípio poderia executar 1018 FLOPS em uma versão ampliada) usando a tecnologia 2016. O problema que realmente me surpreendeu foi que o cálculo real não era o fator limitante em termos de tempo e consumo de energia, mas estava movendo dados para frente e para trás entre processadores, níveis de cache, memória principal e armazenamento e energia necessário para manter o estado nos chips DRAM.

A computação orientada a memória soluciona esses problemas, recolhendo a hierarquia de armazenamento em memória e substituindo, na medida do possível, DRAM, armazenamento em estado sólido e discos rígidos por memória não volátil, de alta densidade e

baixo custo (consulte a Figura 1). Isso abrirá o gargalo de von Neumann que assolou a computação desde o advento dos primeiros computadores de programas armazenados. A ideia básica é que exista um núcleo muito grande (potencialmente muitos petabytes ou mais) de memória compartilhada mantendo todo o conjunto de dados de interesse, cercado pelo processamento de tipos apropriados. Os avanços na fotônica descritos abaixo permitem que esses pools de memória escalem economicamente de gigabits a petabytes, de miliwatts a megawatts e de dezenas a centenas de milhares de dispositivos computacionais. A memória se torna central e não volátil, enquanto a computação se torna periférica e efêmera, organizada para avançar o estado da memória e, em seguida, tão rapidamente desativada.

Originalmente, as operações de E / S eram usadas para isso, obtendo e entrando dados do computador a partir de periféricos que estavam realmente localizados na periferia do sistema. Porém, com o tempo, sobrecarregamos as operações de E / S com funções relacionadas à memória: comunicação entre processos, compartilhamento de memória e cluster. A computação orientada a memória usa a malha de memória para essas funções e retorna a E / S de volta à sua função original. Todos os outros usos são substituídos pelo ponteiro que passa na região da memória compartilhada. Para que esse conceito seja viável, a memória não volátil deve ser significativamente menos cara por bit para ser acessível e consumir muito menos energia do que a DRAM, mas não precisa ser tão rápida quanto os dados forem endereçáveis por byte. As simplificações do hardware e do sistema operacional ainda podem tornar o desempenho no nível do sistema superior a uma arquitetura hierárquica convencional, na qual todos os dados são coletados de onde estão armazenados e enviados para um local central para processamento, pelo menos para muitos tipos de computação intensiva em dados.

Interconexão / tecido fotônico

Depois que o gargalo de memória é aberto, somos imediatamente confrontados com a necessidade de aumentar a velocidade e reduzir a latência na interconexão. James Meindl previu que a "tirania dos interconectores" acabaria se tornando o principal fator limitante para a eletrônica.6 A interconexão fotônica não era considerada viável para computadores, principalmente por causa da despesa de equipamentos de comunicação que estavam disponíveis. No entanto, as vantagens tecnológicas dos fótons sobre os elétrons em termos de dissipação de energia, integridade do sinal, largura de banda por canal da multiplexação por divisão de comprimento de onda rompendo os limites de entrada / saída de chips e latência diminuída devido à velocidade intrínseca e à eliminação de repetidores, despertaram interesse e desenvolvimento renovados.

Os principais problemas a serem resolvidos foram a eficiência energética da conversão eletrônica em fotônica (e o oposto), acoplando a luz do laser dentro e fora da matriz e, o mais importante, o custo de fabricação e montagem. Todas essas questões foram abordadas nas fundições CMOS padrão para produzir "fotônica de silício", que incluem guias de onda de silicone microfabricados, acopladores de grade, moduladores e detectores integrados diretamente à eletrônica de controle analógico / digital. A camada fotônica e o plano de controle eletrônico podem ser fabricados em matrizes separadas e, em seguida, flip-chip ligados a processadores, aceleradores ou outros tipos de chips e pacotes para fornecer a conectividade dentro da estrutura do computador. A interconexão fotônica é agora uma rápida disciplina de engenharia crescente e avançada que está nos estágios avançados de desenvolvimento e comercialização inicial. O potencial para melhorias drásticas no desempenho da interconexão eletrônica sempre foi óbvio, mas o principal obstáculo à adoção sempre foi o custo, que finalmente está sendo resolvido por meio da integração.

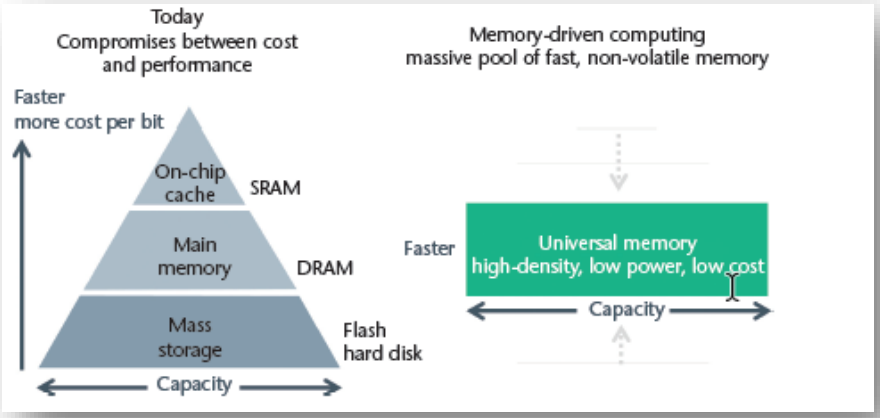


Figura 1. Ilustrações esquemáticas da hierarquia tradicional de memória e armazenamento que define os computadores atualmente (à esquerda) e a estrutura achatada possível com a introdução de memória não volátil, barata, de alta densidade e baixa potência (à direita). Atualmente, os computadores podem ter uma hierarquia de tipos de memória e armazenamento com até 10 camadas de profundidade definidas pela economia; os usuários compram o máximo de memória rápida possível e fazem backup com armazenamento barato para manter o que não cabe na memória. No entanto, o resultado é que até 90% do trabalho de um computador pode embaralhar dados entre camadas, o gargalo de von Neumann, em vez de executar os cálculos exigidos pelos usuários. A memória não volátil pode substituir muitas das camadas da hierarquia por uma única tecnologia, mantendo todo o conjunto de dados para um problema e eliminando o tempo e a energia desperdiçados no embaralhamento de dados.

Aceleradores

A introdução de uma plataforma de computação orientada a memória ativada pela interconexão fotônica é uma grande mudança na arquitetura do computador, mas é um evento único que pode não ser replicado por décadas. Uma vez que as vantagens de uma nova arquitetura são percebidas, o que pode ser feito para melhorar continuamente o desempenho e a eficiência do hardware do computador? Um cenário possível é que aceleradores especializados sejam construídos para abordar classes de computação ou mesmo aplicativos específicos, para que os computadores possam ser mais facilmente customizados para problemas específicos. Geralmente, agora pensamos em aceleradores para computadores como GPUs ou FPGAs, e certamente esses tipos de processadores oferecem vantagens significativas em relação às CPUs de uso geral para certos tipos de computação. No entanto, eles ainda são circuitos CMOS convencionais que precisam ser programados, o que limita sua velocidade e eficiência energética.

Em muitos sistemas eletrônicos, os ASICs são projetados e construídos para executar uma tarefa específica que pode substituir um FPGA por uma maior velocidade e eficiência e menor custo total. Isso só é economicamente viável se o mercado desses chips for grande o suficiente para que os custos de engenharia não recorrentes associados ao projeto e fabricação dos ASICs possam ser suficientemente amortizados. Como alternativa, nas duas últimas décadas, muitas das tarefas anteriormente executadas pelos ASICs em um computador foram integradas às CPUs porque eles tinham um número cada vez maior de transistores para desempenhar mais funções. No entanto, com o fim da lei de Moore, tarefas altamente repetitivas que dominam certas aplicações podem ser executadas usando ASICs que podem se comunicar através de uma tela fotônica como um par com um conjunto de processadores heterogêneos e outros aceleradores (veja a Figura 2). Até mesmo blocos significativos de código podem ser traduzidos em hardware usando um designer de chip automatizado como o

PICO (entrada e saída do programa). Assim, os usuários finais podem obter melhorias significativas e contínuas no desempenho de suas aplicações, através de maior personalização e disponibilidade de novos aceleradores de uma plataforma de computação orientada a memória com uma malha aberta.

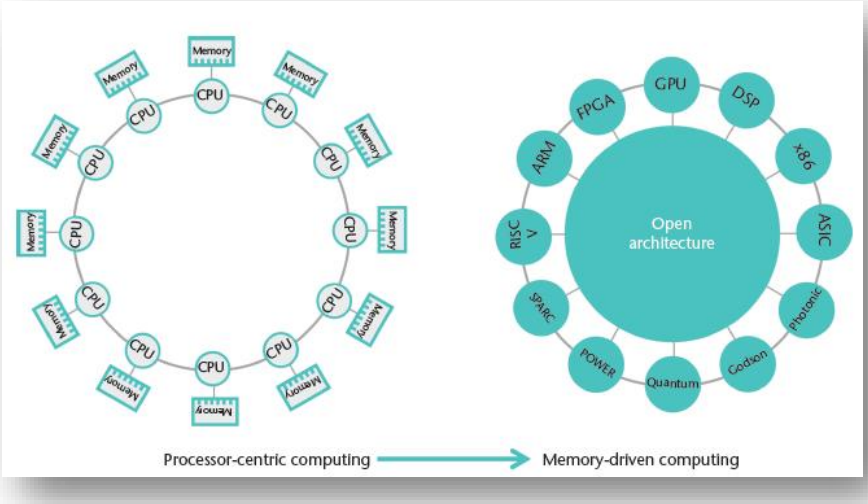


Figura 2. Uma visão alternativa da comparação entre computação centrada no processador (esquerda) e orientada a memória (direita). Uma tela fotônica aberta deve ser capaz de conectar e suportar um ambiente heterogêneo com todos os tipos de processadores, aceleradores ou módulos de memória criados por qualquer fornecedor. Usando uma plataforma comum, os computadores podem ser personalizados de acordo com as necessidades do usuário, selecionando um conjunto apropriado de tecnologias, sejam componentes padrão para computação de alto desempenho ou chips neuromórficos especializados para mecanismos de pesquisa. A natureza aberta do tecido também pode inflamar a inovação, permitindo que qualquer entidade acadêmica, governamental ou comercial que possa projetar um chip seja fabricada em uma fundição e participe do ecossistema de computação.

Uma vez reconhecido o potencial de usar ASICs como aceleradores, agora podemos pensar em termos ainda mais amplos sobre quais novos dispositivos e circuitos poderíamos construir. Um exemplo disso é um "mecanismo de produto de ponto" de multiplicação de matriz de vetores analógico baseado em uma matriz de barras cruzadas de memristores ou comutadores resistivos não voláteis com uma faixa contínua de estados de resistência, que seriam os mesmos dispositivos básicos usados na memória não volátil descrita mais cedo. Este é um exemplo de utilização da física de um sistema, neste caso as leis de Kirchoff, para realizar um cálculo importante. Os elementos em uma matriz positiva com valor real podem ser representados pelas condutâncias celulares correspondentes em uma barra transversal. Se um conjunto de tensões que corresponde a um vetor de entrada for aplicado às linhas da barra transversal, as correntes de saída nas colunas representam o produto escalar do vetor com a matriz. O tempo necessário para executar o cálculo é essencialmente uma única etapa de tempo, independentemente do tamanho da matriz, em vez de O (n2), onde n é o comprimento do vetor, necessário para executar o produto escalar matematicamente. Isso fornece melhorias de ordem de grandeza no tempo e no gasto de energia para essa operação específica, mesmo quando comparado com um CMOS ASIC puro, desde que o aplicativo corresponda à precisão de bits do cálculo analógico.

A aceleração da multiplicação de matrizes vetoriais beneficia uma variedade de aplicações, principalmente processamento de sinais (transformadas discretas de Fourier), processamento de imagens (convoluções, filtragem) e treinamento e inferência de redes

neurais. Os benefícios de desempenho nesses casos surgem da realização de cálculos repetitivos o mais próximo possível do local da memória armazenada. Esses novos circuitos e arquiteturas para computação podem aproveitar ainda mais os dispositivos emergentes de baixa potência e os benefícios combinados pode levar a mais de cinco melhorias de ordem de grandeza na taxa de transferência de processamento dividida por potência (ops / s / Watt) para aplicativos como redes neurais convolucionais. Embora os anúncios do nascimento da inteligência artificial tenham sido quase tão numerosos (e incorretas) como as previsões para a morte da lei de Moore, vários desempenhos inovadores recentemente alcançados com redes de aprendizado profundo prenunciam uma gama diversificada de aplicativos abordados não mais por sub-rotinas de software escritas por programadores, mas por redes neurais artificiais que aprendem diretamente através da exposição a dados.

Inspiração

Além de um acelerador relativamente simples, como o mecanismo de produto escalar, existem outras funções que poderiam ser expressas em um ASIC de tecnologia mista (como dispositivos CMOS + analógicos / digitais que executam cálculos por meio de um processo físico)? Podemos buscar mais inspiração entendendo melhor os tipos de primitivas e algoritmos computacionais no cérebro. Embora nossa compreensão do cérebro hoje seja limitada, sabemos o suficiente agora para projetar e construir circuitos que podem acelerar certas tarefas computacionais. À medida que aprendemos mais sobre como os cérebros se comunicam e processam as informações, devemos aproveitar esse entendimento para criar um novo e, esperamos, exponencial caminho de crescimento para a computação. Até redes neurais relativamente simples são capazes de resolver problemas complexos se tiverem sido treinadas com um conjunto de dados grande o suficiente. Seremos capazes de implementar tarefas relativamente sofisticadas, como reconhecer uma ameaça em potencial de algo nunca visto antes para proteger computadores e datacenters de uma primeira instância de um vírus ou outro ataque? Que tal ser capaz de generalizar da experiência para reconhecer quando a solução para um problema que foi resolvido anteriormente pode ser aplicada a uma nova consulta ou encaminhar um problema específico para uma máquina ou pessoa diferente que provavelmente já sabe a resposta? Nesse estágio, essas são apenas especulações interessantes, mas podem se tornar assuntos legítimos para estudo na ciência da computação, à medida que os neurofisiologistas aprendem mais sobre os circuitos e os psicólogos aprendem mais sobre algoritmos no cérebro.

O fim da lei de Moore pode ser a melhor coisa que aconteceu com a computação em décadas. Isso forçará aqueles que constroem computadores a serem mais criativos, a explorar novos domínios da ciência da computação e a experimentar diferentes arquiteturas, circuitos, conceitos de dispositivos, física e materiais, e a procurar novos paradigmas computacionais da biologia e de outros lugares. O objetivo não é replicar o cérebro ou duplicar exatamente qualquer uma de suas funções, mas encontrar inspiração para novas maneiras de resolver problemas com um computador e utilizá-los em sistemas otimizados para os aplicativos que precisam deles. A ciência e a engenharia podem finalmente estar prontas para enfrentar as aspirações que Turing e von Neumann tiveram pelos computadores na década de 1940, ou o florescimento da criatividade pode nos levar a direções completamente novas e imprevistas. Haverá uma nova tecnologia que permita melhorias exponenciais de desempenho nas próximas décadas? Não sabemos ao certo, e é isso que torna o futuro da computação tão emocionante.

Computing beyond the End of Moore’s Law: Is it really the end, and what are the alternatives?

Fonte: https://www.researchgate.net/publication/288855758_Computing_beyond_Moore's_Law

Resumo

O fim da linha da Lei de Moore está agora à vista, pois os sistemas de fotolitografia estão em ritmo de atingir a escala atômica até o final da próxima década. Desde 1965, a indústria eletrônica mundial passou a depender da escala rápida, previsível e barata do desempenho da computação, que foi possibilitada pela escala da tecnologia. O fim do escalonamento de tecnologia convencional afeta todos os tipos de tecnologias de computação que dependem de melhorias no custo, na eficiência energética e na capacidade de armazenamento - da computação em larga escala aos menores dispositivos eletrônicos de consumo. Este artigo descreve os limites da tecnologia microeletrônica semicondutora existente no nível do dispositivo e seu impacto no nível do sistema. Em seguida, examina alternativas para futuras tecnologias de computação e propõe uma abordagem para avaliar sua viabilidade. Por fim, oferecemos em nossa conclusão um resumo das principais implicações do fim da Lei de Moore e uma perspectiva sobre um roteiro de pesquisa para enfrentar esse desafio.

Introdução

Em 1965, Gordon Moore observou notoriamente que o número de componentes em um circuito integrado vinha dobrando a cada ano, em média, desde a introdução dessa tecnologia em 1959 [1] (embora isso tenha moderado a dobrar aproximadamente a cada 18 meses hoje). Ele previu que essa tendência, impulsionada por considerações econômicas de custo e rendimento, continuaria por pelo menos uma década. Ele também observou que "diminuir as dimensões de uma estrutura integrada torna possível operar a estrutura em velocidade mais alta para a mesma potência por unidade de área" - uma observação que foi formalizada em quase uma década depois por Robert Dennard da IBM como Dennard Scaling. Esse comportamento de dimensionamento que se reforça mutuamente (tamanho do recurso, frequência e potência) significava que a funcionalidade do chip melhoraria exponencialmente com o tempo a um custo aproximadamente constante por geração, e Moore previu que isso, por sua vez, levaria a uma cornucópia de benefícios sociais que fluiriam da microeletrônica semicondutora tecnologia. Os efeitos fortuitos de escala que Moore previu realmente se mantiveram verdadeiros e duraram 40 anos a mais do que ele previra inicialmente. No entanto, essas regras de escala de tecnologia estão enfrentando desafios, pois a escala de Dennard chegou ao fim em 2004 - levando a uma crise de eficiência de energia para a lógica do CMOS. Mas a escala de tecnologia tradicional será confrontada com um desafio ainda mais fundamental na próxima década.

Dentro dessa década, o processo mágico de escala que Moore descreveu terminará quando a capacidade de litografia 2D se aproximar da escala atômica. Esse limite de escala afetará não apenas a tecnologia tradicional de computação digital, mas também todas as outras tecnologias microeletrônicas convencionais. Atualmente, não existe uma tecnologia sucessora óbvia para a lógica onipresente de Complementary Metal Oxide Semiconductor (CMOS). Três opções básicas para o progresso contínuo da computação nos ocorrem: (1) novos dispositivos, (2) novas arquiteturas (com ou sem novos dispositivos) e (3) novos paradigmas de computação. Pode-se esperar uma exploração e inovação substanciais em cada uma dessas áreas.

No curto prazo, provavelmente haverá ênfase no desenvolvimento de dispositivos baseados em CMOS que se estendam para a terceira dimensão, ou vertical, e para melhorias na tecnologia de materiais. É provável que elas co-evoluam com novas abordagens arquiteturais que melhor adaptam a capacidade de computação a problemas específicos de computação, impulsionados principalmente por grandes forças econômicas associadas ao mercado global de TI de US $ 4T / ano. A longo prazo, é provável uma transição para novas classes de dispositivos e o surgimento de sistemas práticos com base em novas abordagens

de computação. Para serem eficazes para atender às necessidades e expectativas da sociedade em um contexto amplo, esses novos dispositivos e paradigmas de computação precisarão ser economicamente manufaturáveis em escala e também precisarão fornecer um caminho de melhoria exponencial. Isso pode exigir uma mudança tecnológica substancial análoga à experimentada na transição da tecnologia de tubo de vácuo para tecnologia de semicondutores.

Essa transição exigirá esforço em uma escala de tempo decadal; portanto, independentemente de o roteiro de semicondutores ter 10 ou 20 anos de vitalidade restante, é importante estabelecer as bases estratégicas para a mudança agora. É crucial se preparar para essa transição, apoiando pesquisas e desenvolvimento relevantes, e nossa intenção é contribuir com uma perspectiva estratégica em relação a essa necessidade.

É realmente o fim?

Longe de uma lei, a observação de Moore foi uma teoria econômica impulsionada pelo Technology Scaling - a constante melhoria nos processos de fotolitografia para diminuir o tamanho dos componentes nos chips. De fato, o dimensionamento da tecnologia é o fator subjacente e existe muita preocupação de que um dos componentes do dimensionamento da tecnologia (melhorias em 2D na litografia de silício) esteja se aproximando dos limites fundamentais até o final da próxima década. Vários ataques à escala de tecnologia convencional para eletrônicos digitais ao longo dos 50 anos de história da Lei de Moore desafiaram a abordagem convencional de melhoria de desempenho.

Por exemplo, antes dos anos 90, a maior parte da lógica digital era baseada na lógica de transistor bipolar. Na década de 1980, a densidade de energia dessa tecnologia se tornou impraticável. A incapacidade de continuar a escalar a lógica bipolar (TTL, ECL etc.) levou a uma transição geral para a tecnologia lógica CMOS (Complementary Metal Oxide Semiconductor), mais eficiente, que possibilitou mais duas décadas de escalabilidade tecnológica. Mais recentemente, em 2004, o dimensionamento de Dennard [2] (a capacidade de reduzir tensões operacionais de dispositivos e frequências de clock de escala a cada geração a taxas exponenciais) começou a falhar, e a indústria de computação estava novamente em uma crise de densidade de potência, semelhante à crise bipolar. Sem crescimento no desempenho por núcleo de processamento, os arquitetos de computadores responderam ao escalonamento exponencial do número de núcleos por chip (tecnologia multicore) para continuar o escalonamento da tecnologia por mais uma década. Hoje, a eficiência energética dos portões lógicos continua a aumentar (embora a uma taxa mais lenta), a capacidade de carga de dados e a eficiência dos fios não estão melhorando substancialmente.

Enquanto os sistemas de programação atuais são projetados para conservar operações lógicas às custas da movimentação de dados, o aumento do custo da movimentação de dados em relação às operações lógicas pressagia uma mudança dos paradigmas de programação centralizados na computação para paradigmas mais centrais. O ponto principal é que, apesar da falha de vários mecanismos físicos subjacentes à tecnologia ao longo de 50 anos de história da lei de Moore., A comunidade já havia encontrado novas abordagens para continuar esse dimensionamento. Um pesquisador brincou: "Prevejo que a Lei de Moore nunca terminará - assim, só estarei errado uma vez!"

Nossa visão é que o escalonamento da tecnologia está agora em sério risco, porque os limites do escalonamento da litografia 2D são fundamentais e porque não existe uma tecnologia sucessora óbvia. Um átomo de silício tem aproximadamente meio nanômetro de diâmetro em material semicondutor. Na atual taxa de aprimoramento, os sistemas de fotolitografia serão capazes de criar recursos de transistor na escala de poucos átomos usando a tecnologia 5nm em 2022-2024 [3]. Esse tamanho de recurso corresponde a uma dúzia ou menos de átomos de Si nos recursos críticos do dispositivo e, portanto, a tecnologia será um limite prático para controlar a carga no sentido clássico. Ir além exigiria a engenharia desses dispositivos em um regime no qual os efeitos da mecânica quântica dominariam, por exemplo,

tunelamento de elétrons através do óxido da porta, o que aumentará as perdas de energia devido à corrente de fuga. Embora seja tecnologicamente viável atingir de 3 a 5 nm em 2022 usando o Extreme Ultraviolet (EUV), a taxa real de adoção de tamanhos de recursos menores é mais impulsionada pela economia e pelo retorno do investimento (ROI) por meio de melhorias de desempenho do que pela viabilidade tecnológica. Além disso, o custo cada vez maior dos métodos litográficos pode tornar sua produção economicamente inviável.

Nesse momento, novas melhorias na litografia não resultarão mais em tamanhos menores de recursos ou desempenho do dispositivo suficientemente melhor para justificar o aumento dos custos. Portanto, o fim da Lei de Moore é realmente o fim de uma escala bidimensional útil da fotolitografia. As restrições impostas pela física fundamental dos dispositivos e o aumento dos custos de fabricação da produção de transistores menores (que são dominados pela litografia) são limites iminentes. O que se tornará insuperável primeiro não é claro. No entanto, concluímos que novas melhorias na densidade do circuito planar se tornarão inviáveis através de um desses mecanismos. Como resultado, o setor de computação não poderá mais ampliar a capacidade de computação usando a abordagem básica que funcionou tão bem desde 1965.

O fim da Lei de Moore afetará todos os dispositivos que dependem do tamanho reduzido do recurso para progredir. Isso inclui dispositivos de processamento e armazenamento [4]. Para aumentar ainda mais a densidade do circuito ou do armazenamento, será necessário construir na terceira dimensão, que exigirá uma tecnologia que suporte o ganho do sinal e reduz a energia consumida pelo movimento dos dados. A resistência intrínseca do material usado pelos “fios” (conexões elétricas de alguma forma) limitará qualquer solução envolvendo elétrons. O principal problema é que a energia consumida para transmitir um bit é proporcional à distância que ele deve percorrer devido à resistência e capacitância do metal usado para conduzir os elétrons que representam o bit [11]. O cobre é um condutor tão bom quanto o esperado para um material comum à temperatura ambiente. Independentemente disso, a movimentação de dados ainda dominará as perdas de energia e restringirá a capacidade de desenvolvimento na terceira dimensão.

A sociedade passou a esperar e confiar nos benefícios proporcionados pela Lei de Moore. O fim da Lei de Moore apresentará desafios de empacotamento e desempenho para todos os tipos de dispositivos eletrônicos de consumo que dependem de melhorias no custo e na eficiência energética para extrair mais funcionalidade de um dispositivo com capacidade limitada de bateria. Por exemplo, a capacidade de tornar-se um telefone inteligente "mais inteligente" será comprometida se o setor não conseguir empacotar mais dispositivos em um espaço menor. A questão mais profunda apresentada por essas mudanças é a ameaça ao crescimento econômico futuro da indústria de computação dos EUA. A escala da lei de Moore transformou a computação em uma tecnologia difundida para o consumidor. À medida que a tecnologia de computação se tornou mais poderosa, ela foi usada de maneira cada vez maior, e o mercado cresceu exponencialmente como resultado. Concluímos que o fim da escala da lei de Moore ameaça causar estagnação na inovação de produtos, o que, por sua vez, pode ter um impacto econômico negativo significativo.

Quais são as alternativas e como avaliamos sua viabilidade?

Um relatório recente da Atividade de Projetos de Pesquisa Avançada em Inteligência (IARPA), intitulado Uma Análise Inicial das Tecnologias Computacionais Alternativas para a Comunidade de Inteligência, postula que, diante desses desafios iminentes, pode ser necessário considerar uma visão mais ampla do que constitui computação. Além disso, conclui que o processo deve começar considerando quatro modelos computacionais básicos:

1. **Computação digital clássica (CDC)**: o CDC inclui todas as formas de eletrônica digital binária que formam a base para toda a indústria de computação e eletrônica de consumo.
2. **Computação analógica (CA)**: dispositivos não binários que implementam a computação através de princípios físicos diretos.

3. **Computação neuro-inspirada (NC)**: Dispositivos baseados nos princípios de operação do cérebro e na computação neuronal geral.
4. **Computação quântica (QC)**: O entrelaçamento quântico poderia, em teoria, ser usado para resolver alguns problemas com complexidade combinatória através da seleção do estado desejado a partir de uma superposição de todas as respostas possíveis para um problema.

Em uma visão importante, o relatório recomenda que seja necessário distinguir entre novos paradigmas de computação e novas implementações tecnológicas de paradigmas existentes. Em particular, o relatório faz as seguintes observações;

- **Computação analógica**: a CA pode ser mais simples do que algumas aproximações digitais, mas não se presta à computação de uso geral, porque a forma do dispositivo é especializada em computação. A precisão computacional é problemática de manter e pode ser sensível ao seu ambiente.
- **Neuromórfico / de inspiração biológica**: Os computadores digitais são bons no cálculo determinístico / algorítmico, mas são pobres em tarefas simples de raciocínio e reconhecimento. Os dispositivos de computação inspirados em neuro demonstraram ser inerentemente resistentes e muito bons em problemas que o CDC não é. Existem muitas oportunidades inexploradas para esses modelos computacionais, mas ainda não se sabe muito sobre como o cérebro realmente calcula.
- **Processamento quântico de informações**: O processamento quântico de informações pode, em teoria, permitir a solução eficiente de alguns problemas combinatórios e NP-Hard (problemas que não podem ser resolvidos em tempo polinomial usando computação digital). Não é um substituto adequado para o CDC em áreas onde essa abordagem se destaca.

Concordamos com a posição do white paper da iARPA [12] de que essas opções de tecnologia criam a possibilidade de abordagens que vão muito além do que as tecnologias CMOS e eletrônica digital tradicionalmente têm realizado de maneira eficaz e, portanto, são dignas de investimento em pesquisa. Mas eles não são adequados como substitutos para a eletrônica digital para tarefas nas quais a computação digital já apresenta um bom desempenho. Portanto, para o restante deste artigo, nos concentramos em novas implementações tecnológicas do modelo CDC, porque este é o modelo com relevância mais imediata para um amplo conjunto de preocupações sociais associadas ao fim da Lei de Moore.

Métricas para avaliar uma alternativa CMOS / CDC

No passado, um concorrente do CMOS / CDC precisaria acompanhar um cronograma incansável de melhorias, no qual o CMOS dobraria seu desempenho a cada 18 meses ou mais e conseguiria alavancar enormes economias de escala. Essa combinação de maior desempenho e economias de escala se mostrou imbatível, exceto em nichos específicos e relativamente estreitos. Com o fim da escala da tecnologia CMOS, essas condições competitivas mudaram. Um concorrente iniciante do CMOS ainda não é aparente, mas podemos aplicar métricas para avaliar a adequação de possíveis concorrentes. Em um artigo de 2006 [5], Shekhar Borkar, da Intel, ofereceu três propriedades necessárias para a substituição do CMOS, às quais John Shalf acrescentou um princípio adicional de manufatura:

1. **Ganho**: a energia necessária para ligar e desligar o estado do dispositivo deve ser menor que a energia que o dispositivo controla.
2. **Imunidade ao sinal de ruído**: o sinal usado deve estar suficientemente acima do nível de ruído de fundo para detectar o sinal.
3. **Escalabilidade**: a tecnologia deve permitir o aumento da densidade e as reduções de energia correspondentes à medida que a tecnologia melhora.
4. **Capacidade de fabricação escalável**: A tecnologia deve ser produzida com um processo capaz de ser implementado em escala industrial.

Possíveis opções de tecnologia pós-CMOS

Esta seção lista as possíveis tecnologias pós-CMOS e descreve a abordagem, os benefícios e os desafios à comercialização de cada uma. Embora não tenhamos realizado uma avaliação detalhada dessas tecnologias, os critérios de Borkar / Shalf e IARPA são usados para avaliar seu potencial.

Empacotamento mais denso: a integração e o empacotamento tridimensionais (3D) estão sendo perseguidos vigorosamente e têm sido bem-sucedidos em dispositivos convencionais para aumentar a densidade lógica e reduzir as distâncias de movimentação de dados. A maioria dos dispositivos de memória enviados hoje possui alguma forma de empilhamento de chips. Exemplos de abordagens para aumentar a densidade de futuros dispositivos usando a integração 3D incluem:

- **Empilhamento de cavacos em 3D usando o Through-Silicon-Via (TSV)**: o empilhamento envolve furos nos cavacos de silício para fornecer conexões elétricas entre as camadas dos cavacos de silício empilhados. Estão disponíveis pilhas de cavacos com até 8 camadas de profundidade, com custos de engenharia mais baixos do que a adição de camadas litograficamente ou deposição epitaxial. O TSV oferece menor largura de banda e conectividade menos eficiente entre as camadas de chip do que adicionar camadas com processos de fotolitografia, mas largura de banda e eficiência muito mais altas do que usar embalagens de chips convencionais.
- **Camadas de metal**: o CMOS é tradicionalmente construído em forma planar 2D, com melhorias modestas em 3D. Os chips modernos têm até 11 camadas de metal. O número de camadas de metal poderia ser melhorado, mas elas fornecem conectividade adicional apenas entre os componentes na superfície 2D.
- **Deposição Epitaxial**: Um problema com as camadas litográficas é que resulta em apenas uma camada do material semicondutor (a camada inferior de silício), que ainda é plana 2D. Para obter transistores mais ativos, é necessário adicionar mais camadas de material semicondutor. A deposição epitaxial envolve um processo de deposição química ou de vapor para adicionar camadas de materiais semicondutores umas sobre as outras. Ainda existem desafios no depósito de camadas ativas de cristal único de alta qualidade, mas há um progresso substancial no estudo de abordagens além do silício padrão, por exemplo, abordagens que usam transferência direta de camadas muito finas de material cristalino a granel usando processos diferentes da deposição epitaxial.

Os principais desafios para dimensionar as camadas litográficas 3D são a tolerância aprimorada a defeitos e o gerenciamento das densidades térmicas e da resistência intrínseca. O empilhamento de tecnologias interessantes, como células de memória não volátil emergentes (MRAM, Memristor etc.), oferece uma oportunidade substancial para permitir camadas litográficas mais profundas (potencialmente algumas melhorias de magnitude). Apesar de empilhar

O 3D reduzirá substancialmente os requisitos de movimentação de dados, que são os principais contribuintes para a densidade térmica, ainda não está claro quanto espaço adicional é oferecido para permitir o empilhamento profundo das camadas lógicas. Enquanto isso, as tecnologias de memória 3D liderarão o caminho na integração 3D. As tecnologias que reduzem as tensões operacionais dos circuitos digitais (que estão paralisadas desde 2004) podem fornecer mais espaço para a construção de circuitos na 3ª dimensão. Muitas abordagens de empilhamento 3D acabam falhando de escala devido aos limites de densidade de energia das camadas lógicas intensivas em empilhamento umas sobre as outras. Qualquer futuro sistema "eletrônico" precisará ter material que reduza a potência de comutação para a lógica e as perdas resistivas para transferência de informações dentro de cada camada lógica constituinte (por exemplo, fios de resistência mais baixa exóticos ou comutadores que exijam tensões muito mais baixas).

Materiais / estruturas avançadas: melhorando o desempenho do transistor (comutador) é uma abordagem para melhorar o desempenho do dispositivo subjacente à maioria dos dispositivos eletrônicos digitais.

- **Transistor de efeito de campo de tunelamento (TFET)**: Os transistores convencionais de efeito de campo (FETs) têm um desempenho do dispositivo limitado pelo balanço de tensão necessário para ativá-lo ou desativá-lo completamente (isso é ganho nos critérios de Borkar). Um TFET usa um material de canal que modula o efeito de tunelamento quântico, em vez da modulação MOSFET clássica de emissão termiônica, para criar um comutador que é mais sensível à voltagem do portão ao ligar / desligar e, portanto, pode operar em uma voltagem mais baixa. Como a dissipação de energia do dispositivo é proporcional à tensão2, a oportunidade de melhorias na eficiência de energia (e, portanto, no futuro da tecnologia) é substancial. Diferentes sistemas de materiais estão sendo investigados, mas a sensibilidade térmica, velocidade, desafios na fabricação confiável (uma vez que sistemas de materiais mais complicados são usados do que atualmente) e outros problemas de escalabilidade desafiam os dispositivos atuais. O desenvolvimento bem-sucedido de dispositivos TFET pode permitir uma ou duas gerações adicionais de melhorias no dimensionamento do desempenho da tecnologia na ausência de melhorias na litografia, mas levará 10 anos para que esses avanços no laboratório possam ser traduzidos para a produção em massa convencional. Outras tecnologias envolvem novos designs de porta para melhorar a sensibilidade do transistor, como os FETs de porta ferroelétrica e outras tecnologias. Todos eles têm desafios semelhantes na fabricação e oportunidades semelhantes para estender a eficiência energética da tecnologia (e, portanto, o desempenho), reduzindo as tensões operacionais.
- **Semicondutores heterogêneos / silício tensionado**: O silício tornou-se o principal material semicondutor para circuitos integrados devido às suas propriedades químicas favoráveis e robustez física. Por exemplo, semicondutores formados a partir de combinações das colunas 3 e 5 da tabela periódica (conhecidas como materiais III-V) e. O arseneto de gálio oferece desempenho muito mais alto, mas está mais sujeito a trincas, forma óxidos de baixa qualidade e exige etapas de processamento químico mais desafiadoras. Em relação ao silício, os desafios de manufatura mantiveram os materiais III-V à margem dos principais eletrônicos digitais. Mais recentemente, houve melhorias drásticas na tecnologia para a integração de materiais III-V (arseneto de gálio, arseneto de alumínio e gálio) como ilhas em substratos de silício a granel, a fim de obter os benefícios manufatureiros, químicos e elétricos do silício, juntamente com os benefícios de desempenho do dispositivo dos materiais III-V incorporados. Isso é realizado esticando o substrato de silício para que seu espaçamento atômico se alinhe ao do material III-V e dopando os materiais III-V com impurezas adicionais, para que o espaçamento atômico se alinhe ao do silício quando este é depositado sobre o silício. substrato. Essa abordagem, ainda em sua infância, continua sendo uma tecnologia de processamento exótica. O arseneto de gálio sofre com o desempenho desequilibrado das portas tipo P e N, que por sua vez afeta sua eficiência nos dispositivos CMOS, mas a combinação com silício usando deposição epitaxial pode superar alguns desses desafios. Essa abordagem pode oferecer uma melhoria de ordem de magnitude em algumas funções do dispositivo, mas até agora muitos materiais III-V não oferecem uma substituição exata para o CMOS.
- **Nanotubos de carbono**: a diferença de banda entre os nanotubos de carbono é muito menor que a do silício, e é necessária muito menos energia para operar dispositivos baseados em nanotubos de carbono. Eles também apresentam menor resistência ao movimento de elétrons, mas isso aumenta a suscetibilidade ao ruído. Os transistores baseados em nanotubos de carbono têm demonstrado níveis de corrente mais altos do que os dispositivos à base de silício, o que, em princípio, lhes permitiria operar com taxas de comutação e eficiência energética muito mais altas. Os dispositivos nanotubos foram demonstrados com propriedades de ganho e rejeição de ruído que são competitivas com os semicondutores clássicos para dispositivos individuais. Apesar

dessas propriedades favoráveis, questões mundanas, incluindo resistência de contato, impedem o progresso nos nanotubos de carbono, e os materiais dielétricos de porta ainda precisam ser totalmente projetados e otimizados para os nanotubos. Além disso, existem nanotubos com uma distribuição de diâmetros e bandgaps. Isso leva a uma variação desafiadora do dispositivo; portanto, a fabricação de nanotubos de alta pureza com diâmetro uniforme permanece um grande desafio. O principal desafio dos nanotubos consiste em encontrar um processo de fabricação escalável, pois os dispositivos atuais exigem a colocação precisa dos tubos individuais para formar transistores e circuitos. Houve avanços dramáticos recentemente nos processos de auto-montagem de circuitos baseados em nanotubos [6], mas ainda há um longo caminho a percorrer antes que um processo competitivo e comercialmente escalável esteja disponível.

- **Grafeno**: Ao contrário dos nanotubos de carbono, o grafeno (uma matriz plana de átomos de carbono) não possui um intervalo de banda, portanto, não é adequado para interruptores digitais que se desligam e têm corrente de fuga muito baixa. A solução mais promissora para isso é transformar o grafeno em fitas muito estreitas. Nesse caso, eles são semelhantes aos nanotubos. Os nanotubos são, na verdade, apenas grafeno enrolado em tubos perfeitamente lisos. O desafio aqui é fabricar nanofitas de grafeno com largura uniforme e bordas atômicas. Nanoribos de grafeno são menos bem desenvolvidos que nanotubos, mas estão surgindo técnicas de síntese inovadoras para fitas que podem ser realmente melhores e mais baratas na produção de largura de fita pura e uniforme do que técnicas para nanotubos.
- **Transdutor piezoelétrico (PET)**: os dispositivos PET utilizam o efeito piezoelétrico no qual um campo elétrico induz um estresse mecânico, alterando o tamanho do material. O uso mais comum das variantes PFET tem sido para sistemas micromecânicos e sensores de força, mas se esses materiais piezoelétricos puderem ser miniaturizados com sucesso, a tecnologia poderá ser usada para formar um relé eletrônico em microescala extremamente rápido (multi-gigahertz) [7] . Essa é uma das muitas abordagens que envolvem abordagens micro-mecânicas para o desenvolvimento de switches de maior desempenho.

Arquitetura de computadores: os avanços tecnológicos descritos até agora se concentram em materiais ou embalagens avançadas. Esta seção fornece exemplos de avanços na arquitetura e software de computadores que podem oferecer um aumento no desempenho.

- **Gerenciamento Avançado de Energia**: As tecnologias atuais de gerenciamento de energia são onipresentes e tipicamente muito grosseiras. O Dynamic Voltage and Frequency Scaling (DVFS) e a aceleração térmica permitem economias diminuindo as frequências do relógio e as tensões mais baixas quando as demandas de computação não exigem desempenho máximo. O DVFS de granulação grossa oferece economia significativa de energia para os dispositivos eletrônicos de consumo atuais (que geralmente estão ociosos), mas são apenas marginalmente benéficos para dispositivos que operam perto de 100% de utilização. Pode haver um potencial adicional de recuperar energia com gerenciamento de energia mais refinado e transições mais rápidas entre os estados de energia e através da direção do software das mudanças de estado de energia.
- **Projeto avançado de circuitos**: foram demonstradas abordagens que permitem que os fios operem em uma tensão mais baixa para conexões de longo curso e depois sejam amplificados de forma eficiente nos pontos finais (há uma perda para a re-amplificação). A comunicação pessoal com Bill Dally indicou uma oportunidade para melhorias de 2x – 3x usando técnicas aprimoradas de projeto de circuitos baseadas nas tecnologias atuais.
- **Lógica Clockless / Domino**: Sabe-se que a distribuição de clock consome uma grande fração da energia do sistema e consigna projetos de circuitos para operar na

velocidade de seu componente mais lento. Projetos práticos e eficazes sem relógio provaram ser ilusórios, mas exemplos recentes mostraram que essa abordagem é muito promissora para reduzir o consumo dinâmico de energia, tanto para aplicações neurmorficas quanto digitais.

- Especialização System-on-Chip: O preceito principal da tecnologia System-on-Chip (SoC) é que o custo do chip é dominado pelos custos de projeto e verificação de componentes de circuitos individuais. Portanto, é economicamente viável adequar os chips para incluir apenas os componentes do circuito que são valiosos para a aplicação, em vez de seguir a prática atual de design de commodities de projetar um chip que atenda às necessidades de uma ampla gama de aplicações. Essa abordagem é prática comum para chips de telefone celular (como o iPhone da Apple, que usa núcleos de processadores embutidos combinados em um design SoC especializado), mas está apenas começando a ser considerado para o design de chips de servidor e HPC.
- Lógica personalizada: matrizes de portas programáveis em campo (FPGAs) e computação reconfigurável prometem melhorar o desempenho criando circuitos personalizados para cada problema, mas são impedidos pela menor eficiência na implementação. Uma implementação típica de FPGA superprovisiona os fios (por exemplo, a maioria dos fios reconfiguráveis disponíveis em um FPGA permanece sem uso [9]), a fim de maximizar o uso de tabelas de pesquisa (LUTs). Os FPGAs oferecem a oportunidade de melhorar o desempenho em relação às arquiteturas convencionais de conjunto de instruções da unidade de processamento de computador (ISAs), mas um design de circuito integrado personalizado (ASIC) específico para aplicativos oferece uma melhoria de dez vezes o desempenho em relação ao design do FPGA do mesmo circuito devido à eliminação de fiação redundante. Além disso, o design de circuitos é muito caro em relação ao design de software e a tecnologia de computação mais reconfigurável requer conhecimentos substanciais em design de hardware para otimizar o desempenho. É possível que o desincentivo econômico de projetar e verificar circuitos personalizados seja superado pela alternativa de não ter nenhuma escala de desempenho.
- Silício escuro: Os exemplos mais extremos de lógica personalizada propõem uma família de projetos personalizados que são implementados em um ASIC e permanecem desativados ("escuros") quando não utilizados. O conceito é gastar mais área de superfície ASIC para circuitos especializados mais eficientes, com a esperança de que o custo da área extra seja compensado pelos benefícios de desempenho. Essa abordagem permanece neutra em termos de energia, desligando os circuitos especializados quando eles não são necessários. Atualmente, ele é usado com bons resultados em algumas aplicações especializadas em produtos eletrônicos de consumo, mas sua utilidade para computação de uso geral ainda não foi comprovada.
- Operação de tensão de quase-limite (NTV): até agora, a comunidade de computadores mainstream evitou reduções adicionais na tensão do dispositivo porque reduz a imunidade sinal-ruído dos transistores e sujeita os circuitos a variações estatísticas mais amplas no desempenho. Ambos os efeitos apresentam enormes desafios ao desenvolvimento de software e hardware. Do ponto de vista do software, o desempenho não determinístico de circuitos individuais (e, portanto, elementos de processador individuais) tornaria insustentáveis as abordagens síncronas em massa convencionais para escalar o desempenho da computação paralela - exigindo uma mudança em direção a modelos de execução de software inteiramente assíncronos e reformulações correspondentes de algoritmos e infraestrutura. Os desenvolvedores de aplicativos e algoritmos precisariam reescrever substancialmente o software para acomodar essa heterogeneidade imprevisível de desempenho. No hardware, o aumento da falta de confiabilidade exigiria uma detecção mais abrangente de erros e a infraestrutura de software correspondente para responder a esses erros, e o custo dessa detecção não é conhecido. A abordagem também reduziria substancialmente as frequências de clock, colocando mais pressão no paralelismo para obter melhorias de desempenho (um desafio já assustador para o software). A oportunidade oferecida pela

operação do circuito NTV é o potencial de reduzir as tensões operacionais e, portanto, aumentar a eficiência energética dos dispositivos (e, portanto, o desempenho e a escalabilidade utilizáveis) em uma ordem de magnitude. Continua sendo uma área de pesquisa ativa [8] para determinar se os desafios de software colocados pela confiabilidade, heterogeneidade de desempenho e paralelismo aumentado prejudicam a potencial melhoria bruta de desempenho oferecida.

A redução da resistência é realizada na maioria dos circuitos integrados modernos, usando fios à base de cobre para interconectar circuitos. O cobre é um condutor particularmente bom e, à temperatura ambiente, existem poucas opções que oferecem menor resistência elétrica, além de:

- Supercondutor: O supercondutor pode ser um caminho a seguir para melhorar o desempenho do sistema HPC, mas forçará uma saída do mainstream, pois é improvável que o resfriamento seja prático para dispositivos de consumo. Mesmo os supercondutores de alta temperatura à base de Cuprate têm requisitos de resfriamento e blindagem magnética impraticáveis para aplicações de consumo. A viabilidade de eletrônicos resfriados criogênicamente em telefones ou laptops convencionais é duvidosa. Existe um caminho tecnológico para usar eletrônicos resfriados a crio para estender o desempenho da HPC, mas isso implicaria um afastamento do caminho de alavancar a tecnologia de componentes de commodities. Isso pode ter repercussões significativas na competitividade doméstica dos EUA de HPC e na acessibilidade dos sistemas de HPC, que depende de nossa capacidade de alavancar o mercado de eletrônicos de commodities.
- Metais Cristalinos: O cobre é amplamente utilizado para interconectar camadas de designs de chips (através de fios). O cobre é um condutor muito bom, mas em uma configuração policristalina típica, os elétrons ainda se espalham para fora dos limites entre os grãos cristalinos vizinhos. A condutividade das camadas de metal pode ser melhorada em até um fator 5, criando tamanhos de grãos maiores. As técnicas para criar grãos de cristal maiores em um processo de fabricação de chips escalável ainda não são bem conhecidas (ou talvez simplesmente não estejam sendo compartilhadas por motivos de propriedade).

Altere como os bits são armazenados e transformados: Outros sistemas avançados de materiais fornecem um caminho crítico para melhorar o desempenho dos sistemas eletrônicos. Os materiais apresentados acima são capazes de melhorar o desempenho de dispositivos usando arquiteturas e modelos computacionais familiares de computação digital. Os dispositivos discutidos abaixo também podem melhorar a eletrônica digital, mas exigem um afastamento dos princípios da computação em estoque.

- Spintronics: A computação e a comunicação de informações através da manipulação de domínios magnéticos são mais baixas em custos de energia do que mover elétrons a tal ponto que é quase inconseqüente ao consumo geral de energia. A SEMATECH declarou que os materiais de spin prometem os conceitos de funcionalidade dupla (lógica + memória) e novo circuito (SRAM). Para aplicativos que envolvem tecnologias de memória, há pouco impacto nos paradigmas padrão de computação, mas o uso mais amplo de dispositivos spintrônicos como aplicativos de computação de uso geral (substituições completas do CMOS) exigiria um modelo de computação adiabático ou reversível. Tais modelos de computação podem ser altamente restritivos e perturbariam fundamentalmente nosso modelo atual de computação.
- Isoladores topológicos: Comparados aos fios convencionais, os estados de energia confinada em 2D podem oferecer transporte e armazenamento de informações mais eficientes (margem de ruído mais alta), mas a abordagem adequada para implementar a lógica é incerta. Por exemplo, isoladores topológicos oferecem um caminho possível

para algoritmos de análise de imagem 2D usando efeito foto-galvânico para programar o estado inicial de qubits embutidos no isolador topológico. De acordo com a SEMATECH, novos semicondutores com propriedades únicas, como os semicondutores 2D, estão sendo considerados com cuidado pela indústria eletrônica.

- Nanofotônica: Os principais desafios no uso da nanofotônica em escala de sub-comprimento de onda como substituto das tecnologias de computação / transistor é o baixo ganho de "transistores" ópticos disponíveis e o grande tamanho dos comprimentos de onda ópticos em comparação com as atuais escalas fotolitográficas realizáveis. O benefício mais óbvio da tecnologia fotônica é a comunicação escalável. Para comunicações (substituição de fios), a fotônica tem o benefício de ter custos de energia quase independentes da distância que os dados percorrem, enquanto os fios elétricos padrão têm um forte custo de energia dependente da distância. Portanto, a tecnologia fotônica supera a limitação fundamental da resistência do fio. Infelizmente, atualmente, a energia necessária para acender o laser para enviar informações por uma conexão fotônica é muito superior ao custo do fio, mas diminui constantemente ao longo do tempo [10]. A fotônica terá um papel essencial na superação dos limites dos fios e na disparidade entre os custos de comunicação dentro e fora do chip [11]. O desenvolvimento de um transistor óptico de alto ganho eficaz permitiria que as nanofotônicas também fossem competitivas como substituto do CMOS para computação, mas a tecnologia para um comutador óptico controlado de alto desempenho requer desenvolvimento adicional.
- Computação Química / Biológica: O potencial dos dispositivos de base biológica é que ele tome como inspiração as máquinas mais complexas conhecidas no cérebro de animais terrestres. Os principais desafios para dispositivos de computação com base biológica incluem baixo ganho, baixo sinal para ruído e condições operacionais exóticas. A busca continua por um mecanismo de comutação química que ofereça ganho e rejeição de ruído suficientes para competir com o silício. Existem bons candidatos, mas, além de exigir ambientes operacionais exóticos, os exemplos atuais são difíceis de dimensionar.

As opções para estender a escala tecnológica da eletrônica digital além da Lei de Moore estão resumidas na Tabela 1. Esta revisão não é de forma alguma abrangente, mas fornece uma visão geral de algumas das opções mais discutidas. Como nenhuma das opções é claramente superior em todos os aspectos, é provável que uma ou mais dessas opções cheguem ao uso principal por meio da integração com plataformas convencionais de silício / CMOS. De fato, o empilhamento de chips já está permitindo que a tecnologia fotônica seja empilhada diretamente nos circuitos lógicos e de memória convencionais de silício.

Table 1: Summary of technology options for digital electronics.

Improvement Class	Technology	Timescale	Complexity	Risk	Opportunity
Improving Transistor Performance	Tunnel Field Effect Transistors (TFET)	Mid-Term	Low	Medium	High
	Heterogeneous Semiconductors/Strained Silicon	Mid-Term	Medium	Medium	Medium
	Piezo-Electronic Transistor	Far-Term	High	High	High
	Carbon Nanotubes and Graphene	Far-Term	High	High	High
Computer Architecture Advances	Advanced Energy Management	Near-Term	Medium	Low	Low
	Advanced Circuit Design	Near-Term	High	Low	Medium
	Near Threshold Voltage (NTV) Operation	Near-Term	Medium	High	High
	System on Chip Specialization	Near-Term	Low	Low	Medium
	Custom Accelerators/Dark Silicon	Mid-Term	High	High	High
3D Integration	Chip-Stacking in 3D using Thru-Silicon-Via (TSV)	Near-Term	Medium	Low	Medium
	Metal Layers	Mid-Term	Medium	Medium	Medium
	Epitaxial Deposition	Mid-Term	High	Medium	Medium
Reducing Resistance	Superconducting	Far-Term	High	Medium	High
	Crystalline Metals	Mid-term	unkown	Low	Medium
Other Advanced Materials Systems	Spintronics	Far-Term	Medium	High	High
	Topological Insulators	Far-Term	Medium	High	High
	Nanophotonics	Near/Far-Term	Medium	Medium	High
	Chemical/Biological Computing	Far-Term	High	High	High
	Diamond Films	Far-Term	High	High	Medium

O desafio de longo prazo da tecnologia pós-Lei de Moore exige investimento em ciências básicas, incluindo ciência dos materiais, para estudar materiais de substituição de candidatos e física alternativa de dispositivos para permitir a continuação do dimensionamento da tecnologia. Usando a história do transistor de efeito de campo de aleta de silício (FinFET) como guia, leva cerca de 10 anos para um avanço na física básica de dispositivos para torná-lo um uso convencional. Qualquer nova tecnologia exigirá um longo prazo de entrega e P&D sustentado da ordem de 10 a 20 anos. O vencedor desta corrida tecnológica influenciará não apenas a tecnologia de chips - definirá a direção para toda a indústria de computação.

Conclusão

O fim da Lei de Moore apresenta grandes desafios tecnológicos para a sociedade. De agora até o final da Lei de Moore, o custo de energia da movimentação de dados se tornará um fator técnico e econômico dominante, porque o custo de energia das operações computacionais está melhorando a uma taxa mais rápida do que o custo de energia da transferência de dados para essas operações. Uma adaptação a curto prazo disso será aumentar o uso do paralelismo em software, o que exigirá um grande esforço comercial. Com o tempo, provavelmente será necessário ir ainda mais longe, passando de um modelo de computação centrado em computador para um modelo de datacêntrico.

No curto prazo, espera-se ênfase no desenvolvimento de dispositivos baseados em CMOS que se estendam para a terceira dimensão, ou vertical, e para melhorias na tecnologia de materiais. É provável que elas co-evoluam com novas abordagens arquitetônicas que melhoram a adaptação da capacidade de computação a problemas específicos de computação. Essa evolução será impulsionada principalmente por grandes forças econômicas associadas ao mercado global de TI de US $ 4T / ano, mas investimentos bem direcionados em P&D podem desempenhar um papel importante de nutrição.

A longo prazo, esperamos uma transição para novas classes de dispositivos e o surgimento de sistemas práticos com base em novas abordagens da computação. Para serem eficazes para atender às necessidades e expectativas da sociedade em um contexto amplo, esses novos dispositivos e paradigmas de computação precisarão ser economicamente fabricáveis em escala. Eles também precisarão fornecer um caminho de melhoria exponencial. Isso pode exigir uma mudança tecnológica substancial análoga à experimentada na transição da tecnologia de tubo de vácuo para tecnologia de semicondutores. Espera-se que essa transição ocorra em uma escala de tempo decadal; portanto, se o roteiro do CMOS ainda tem 10 anos ou 20 anos de vitalidade, é importante estabelecer as bases estratégicas para uma grande mudança agora. A indústria deve se preparar para essa transição patrocinando pesquisas e desenvolvimento relevantes e articulando uma visão estratégica abrangente.

Glossário

1. Sistemas microeletromecânicos (MEMS) é a tecnologia de dispositivos microscópicos, particularmente aqueles com partes móveis. Ela funde a escala nano em sistemas nanoeletromecânicos (NEMS) e nanotecnologia. MEMS também são referidos como micromáquinas no Japão, ou microssistemas de tecnologia (MST) na Europa. Os Sistemas microeletromecânicos são sistemas “inteligentes” em miniatura, consistindo em um grande número de dispositivos mecânicos integrados a grandes quantidades de elementos elétricos, sobre um substrato de silício. Os dispositivos mecânicos podem ser de dois tipos: microssensores e microatuadores.

2. O CAGR (Compound Annual Growth Rate), ou taxa de crescimento anual composta, é a taxa de retorno necessária para um investimento crescer de seu saldo inicial para o seu saldo final. Dessa forma, o CAGR é considerado um dos principais indicadores para analisar a viabilidade de um investimento.

3. CMOS Image Sensor

4. AR/VR Augmented Reality/Virtual Reality – Realidade Aumentada/Realidade Virtual