



Patrocinado por



O Futuro da Refrigeração Suplemento



Se preparando para um data center de alta densidade

TEOR

O que fazer com o calor

> Cada watt de potência do data center deveria ser usado duas vezes

Dentro da microfluídica

> Refrigerar dentro de um chip é mais frio e mais interessante do que você imagina

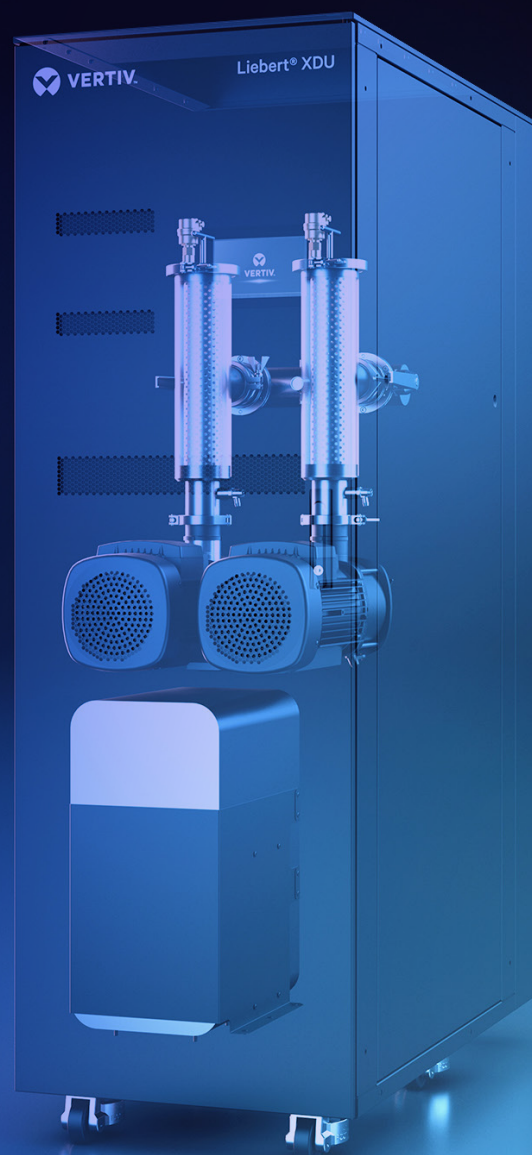
Refrigeração de precisão

> Milhares de jatos minúsculos podem estar chegando em um data center perto de você



Feito para
Densidade.
**Pronto para
Aceleração.**

Gerencie o aumento de calor com
a tecnologia de refrigeração líquida.



Veja como podemos ajudá-lo em [Vertiv.com/GetCool](https://www.vertiv.com/GetCool)

Conteúdo

4. Combinando calor e computação

Cada watt de potência do data center deveria ser usado duas vezes

8. Advertorial: Quando menos não é mais

Operadores de data centers precisarão adotar estratégias de refrigeração híbridas para lidar com as crescentes densidades de TI. O que isso significa?

10. Microfluidica: Refrigerando dentro dos chips

Se você pensa que a imersão é o destino final da refrigeração líquida, repense.

13. Refrigeração de precisão no nível do chip

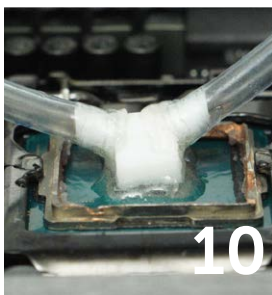
Como milhares de jatos minúsculos podem estar chegando em um data center perto de você

15. Por que a Intel está entrando na área de refrigeração

Como resfriar processadores de 2kW



4



10



13

A Física da Refrigeração

Cedo ou tarde a termodinâmica chega perto de todos nós, e não se pode negociar com estas leis. Os data centers agora usam uma quantidade considerável de energia e geram um nível de calor que impacta seus arredores.

Para lidar com isso, precisamos encontrar a forma mais criativa de trabalhar dentro das regras decretadas pela física.

E essa questão tem implicações no nível macro, para todo o prédio e para as áreas no entorno.

Ela também abre possibilidades no nível micro, até nos processadores e mesmo dentro deles.

Este suplemento mostra como engenheiros podem encontrar um caminho futuro para a última geração de tecnologias para data centers.

Use tudo duas vezes

Toda a energia que usamos nos sistemas de TI se transforma em calor. O calor é um dos resultados inevitáveis e ele é uma mercadoria valiosa.

Então, o que acontecerá se você mudar seu ponto de vista e ver os data centers como uma fonte de calor, e que também fornece computação como um subproduto?

Esse pensamento leva à reutilização do calor e isso traz a possibilidade de sistemas de calefação distritais, onde os data centers canalizam seu ar quente e sua água quente para seus vizinhos.

Se pudéssemos garantir que toda nossa potência computacional fizesse dois trabalhos, poderíamos ter toda a potência que precisássemos –

mas há mais de uma forma para fazer isto.

Entrando no interior dos chips

A medida que os chips se tornam mais complexos e incluem mais transistores, fica cada vez mais difícil resfriá-los. Já há anos, os chips de aquecimento rápido têm sido resfriados por dissipadores térmicos especializados, por sua vez refrigerados por líquidos.

Mas, e se pudéssemos canalizar os fluidos para mais perto dos transistores, dentro dos chips?

Ocorre que os processos que gravam e imprimem componentes eletrônicos minúsculos em um substrato podem ser adaptados para fazer canais microscópicos que transportam refrigerante para o coração do processador.

Formas para ter Chips Mais Frios

A comunidade do silício está encontrando outras técnicas de refrigeração menos invasivas, mas igualmente radicais, para viabilizar chips com TDPs (potência térmica de projeto) ainda maiores.

O programa COOLERCHIPS do governo americano está reunindo diversas abordagens para refrigerar sistemas nesse nível, incluindo esquemas como o JetCool, que dispara jatos de refrigerante dentro de sistemas de refrigeração direta ao chip.

O DOE também atraiu a Intel. A gigante do silício percebeu que seus esquemas para ter chips de silício ainda mais potentes seriam malsucedidos se não houvessem maneiras eficazes para refrigerar esses novos processadores.

A Intel combina dissipadores térmicos gerados por computador e novos materiais.

Todo mundo está buscando técnicas radicais para viabilizar um crescimento contínuo.

Combinando Calor e Computação



Peter Judge
Editor Executivo

Cada watt de potência do data center deveria ser usado duas vezes

A Termodinâmica afirma que toda energia eventualmente se transformará em energia térmica. Se você usar 1 kWh de energia elétrica para rodar um computador, ele emitirá 1 kWh de calor.

Astrid Wynne da TechBuyer explicou assim em um artigo de 2022: "A primeira lei da termodinâmica, também conhecida como a lei da conservação de energia, afirma que a energia não pode ser criada nem destruída, mas ela muda de uma forma para outra. A

energia elétrica que os data centers usam em seus servidores é quase toda convertida em calor. A geração de dados é um subproduto disso, mas é responsável por menos do que 0,1% da conversão."

Para o engenheiro do data center, isso é um problema, pois aquele calor precisa ser removido ou os equipamentos sobreaquecerão.

Para o planejador urbano, esta é uma oportunidade. A comunidade quer esse calor.

Use a sua energia duas vezes

Wynne diz: "O termo 'data center' é um termo impróprio no que diz respeito a energia. Seria mais preciso chamá-los de 'heat centers', ou 'centros de calor' que também podem ser usados para processar e armazenar informações. Inverter o pensamento dessa forma nos permite ser mais criativos em relação a como planejamos e gerenciamos a construção de data centers agora e no futuro."

As comunidades precisam de calor, e a calefação atualmente gera muitas emissões.

No Reino Unido, por exemplo, a calefação responde por aproximadamente 37% das emissões totais. Uma parte desse percentual vem do calor industrial de alta temperatura, mas mais da metade vem simplesmente do aquecimento do ambiente e água quente.

Atualmente, a maior parte desse calor é gerado diretamente de combustíveis fósseis ou por eletricidade, a qual também deriva em grande parte de combustíveis fósseis.

Para atingir emissões net zero, a calefação precisa se descarbonizar, e é por isso que governos estão tentando aos poucos eliminar boilers a gás e buscando por fontes de aquecimento alternativas de baixo carbono. As bombas de calor geotérmicas são uma dessas fontes, mas precisam ser alimentadas por energias renováveis.

O calor produzido pela computação é efetivamente carbono zero, pois é um subproduto da energia usada em outro lugar.

E se o calor é usado em outro lugar, ele provavelmente cancela o ônus da computação sobre a infraestrutura de energia. Você obtém 1 kWh de computação a partir de 1 kWh da energia que já iria se transformar em calor de qualquer maneira.

Não há, teoricamente, praticamente nenhum limite para a demanda por calor residual.

Como observou Mark Bjornsgaard, CEO de Deep Green Energy, em um [podcast](#) da DCD, o total necessário para o calefação predial na Europa é imensamente maior do que a quantidade de energia usada nos data centers.

Por exemplo, o Reino Unido usa ao redor de 434 TWh de energia a cada ano para deixar os prédios aquecidos. O total de energia usado nos data centers é uma minúscula fração dessa quantidade, estimada ao redor de 2 ou 3 TWh por ano.

Isto é 2 a 3 TWh de calor produzidos a cada ano nos data centers. Se todo esse calor pudesse ser usado para calefação, então - até certo ponto, esses data centers teriam provavelmente alcançado net zero no uso operacional de energia. E com mais de 400 TWh de calor sendo ainda necessários, há margem também para alimentar o crescimento projetado para a IA.

Essa é uma hipótese tentadora, mas não tão fácil na prática - ou é?

A dificuldade das redes de calor

Os data centers vêm oferecendo seu calor residual há anos. Mas na verdade, muito pouco tem sido usado com sucesso.

Data centers oferecem um fornecimento de calor previsível pois eles mantêm seus servidores em funcionamento continuamente. Mas o calor é de "baixa qualidade": Ele é morno ao invés de quente, e vem na forma de ar - que é difícil de transportar.

Então, a maioria dos data centers libera seu calor na atmosfera.

Algumas vezes, há redes de calefação distritais que fornecem aquecimento a residências e empresas locais através de uma rede de tubulação. Se o seu data center estiver próximo a uma dessas redes, é uma questão de trazê-la para se conectar ao data center e aumentar o grau de calor.

Você precisa estar no lugar certo para se conectar a uma. "Alguns países estabeleceram ou desenvolveram redes de calefação, mas a maioria não tem uma rede de calefação propriamente dita, então, ocorre de forma fragmentada", Neal Kalita, diretor sênior de potência e energia na NTT, disse à DCD.

É improvável que você encontre uma nos EUA, diz Rolf Brink da consultoria de refrigeração Promersion: "Os Estados Unidos é um ecossistema fundamentalmente diferente. Mas a Europa tem uma densidade populacional mais alta e há mais demanda por calor."

Os países nórdicos têm muitas redes de calefação. A Stockholm Data Parks é um exemplo bem conhecido - um campus de data center na Estocolmo urbana, onde cada data center tem uma conexão à rede de calefação distrital e ele é remunerado por esse calor.

A Alemanha tem algumas redes de calefação e planeja expandi-las. O Reino Unido tem algumas, mas a experiência tem sido difícil até agora.

"Acredito que o desafio dos data centers conectando seu calor em redes de calefação está no quanto aquela rede de calefação estará operacional", diz Kalita.

Qual a qualidade do seu calor?

Outro grande desafio é a qualidade do calor: "Não é o que chamamos de calor de alta qualidade ou calor de alta temperatura", diz Kalita. "Você não pode ter uma rede de calefação completa a partir do calor de um único data center.

Você precisa instalar uma bomba de calor e colocar mais energia para aumentar a temperatura, ou a rede de calefação precisa ter algum tipo de fonte de calor."

A rede de calefação deve estar muito próxima e o calor no ar quente deve ser complementado com outras fontes de

energia, como a queima de biomassa ou o uso de uma bomba de calor para aumentar a temperatura até que possa ser canalizado para o sistema.

Na Suécia, por exemplo, a especialista em computação Cloud&Heat colocou um data center composto de dois containers de 20 pés refrigerados a líquido e equipados com 1600 placas gráficas de alta performance direto em uma instalação de calor e energia combinados (CHP), também chamado de cogeração

O sistema CHP queima biomassa para gerar calor e eletricidade. Os sistemas da Cloud&Heat usam um pouco dessa eletricidade e seu calor residual pode ir diretamente para a corrente de calor de forma que não haja desperdícios.

Calor de baixa qualidade é um problema menor com os novos sistemas de calefação distritais de "quarta geração", os quais se adequaram para lidar com calor de baixa qualidade através de medidas como canos de polietileno que podem manter o calor da água quente.

Introduza a agricultura

Sabe-se, no entanto, que em várias partes do mundo as residências não precisam de aquecimento no verão. A resposta para isso é planejamento, e encontrar outras utilidades. A produção de alimentos tem diversos processos que necessitam de calor, de estufas a fazendas de piscicultura, fermentação e produção de queijos. Mesmo em países com verões quentes, diz Wynne, o calor ainda pode ser usado: "Quando o clima fica mais quente, por exemplo, a energia pode ser direcionada da calefação doméstica para sites de fermentação."

A Green Mountain tem dado água quente de seu data center na Noruega para uma fazenda de piscicultura. O Lefdal Mine subterrâneo em outro lugar na Noruega faz o mesmo, e o Data Center White refrigerado com neve no Japão usa o calor para fazendas de enguias

Valorizando o seu calor

Outro problema é que, como os data centers veem o calor como um produto residual, eles não o valorizam. Redes de calefação são um



negócio, e esperam comprar calor em bases comerciais de um fornecedor comprometido com um fornecimento previsível. Elas querem um PPA (contrato de compra de energia) para o calor.

“Se você entrar em um PPA de calor, então a expectativa é que você forneça consistentemente aquele nível de qualidade do calor e aquele volume de calor”, diz Kalita. “Mas há momentos em que o data center para, ou um cliente sai, ou a instalação precisa ser renovada – e então o calor desaparece.”

“Quando você está comprometido em fornecer calor por 10 ou 20 anos, isso pode restringir a sua flexibilidade.”

Para alguns, essa questão não é intransponível. O consultor de reutilização de calor David Gyulnazaryan disse a um público de data center em Londres que o calor residual é “uma fonte de calor altamente sustentável e confiável” e, se necessário, pode ser “respaldada pela geração de energia no caso de falhas, o que oferece um nível de confiabilidade de 99,996%.”

Pelo lado financeiro, Gyulnazaryan diz “o custo de converter calor de baixa qualidade do data center é de aproximadamente €25/MWh, enquanto o preço do gás natural está ao redor de €200/MWh.

O potencial de economia monetária é enorme para entidades públicas, especialmente quando o fornecimento de gás natural está sendo reduzindo.”

Brink diz que é uma questão de planejamento: “Quando efetivamente pensado durante a fase de construção da instalação do data center, o calor pode ser um grande ativo, com reduções fantásticas em seu TCO.”

“Ao escolher um local para construir um data center, a reutilização do calor normalmente não é ponderada, mas você pode colocar seu data center próximo a uma piscina, ou próximo de alguma coisa que tenha um uso para o calor.”

Ele diz: “A discussão em torno da reutilização do calor está um pouco desequilibrada simplesmente porque se não for colocada no escopo do site, será sem sentido discutir depois – somente porque você não está selecionando o site para aquele fim.”

A regulamentação promove o compartilhamento de calor

O maior problema com os sistemas de calefação distritais pode ser que eles precisam de um investimento enorme e muitos anos para serem implementados. Alguns governos locais e nacionais estão interferindo para encorajar o uso de redes de calefação.

Na região de Amsterdam, na Holanda, os data centers precisam ser projetados para poderem ser conectados às redes de calefação distritais. A Alemanha tem regras

semelhantes em todo o país, que emanam de seu recém aprovado Energy Efficiency Act.

Esta não foi uma questão simples: os primeiros rascunhos da lei exigiam que novas instalações construídas a partir de 2025, reutilizassem 30% do seu calor. A Associação Alemã de Data Centers (GDA) objetou e a lei final diz que os data centers acima de 200 kW precisam apenas reutilizar 20% do seu calor – e não precisam fazê-lo até 2028.

A GDA ainda tem objeções, uma vez que essa exigência limita efetivamente os data centers à locais onde haja sistemas de calefação funcionando.

Benjamin Brake, chefe do departamento de Políticas Digitais e de Dados do governo no Ministério Federal de Negócios Digitais e Transportes (BMDV), reconheceu em conferência da GDA que: “O uso de calor residual como o único critério para escolher uma localização é problemático, especialmente quando não há requisitos nem obrigações para usuários de calor residual.”

No Reino Unido, há ao redor de 14.000 redes de calefação, em sua maioria pequenas, de acordo com a [National Housing Federation](#). O Fuel Poverty Action Group é [crítico](#) do estado das redes de calefação inglesas, alegando que clientes domésticos têm poucos direitos, uma vez que as redes não são regulamentadas. Elas também são vulneráveis a faltas de energia e grandes

aumentos de preço já que a maioria das redes contam com calor de gás natural, cujo preço disparou.

O governo tem um [funil](#) de novos projetos de calefação distrital que somam £3 de libras (3,8 bilhões de dólares). Apenas um dos projetos em Park Royal, com um subsídio de 36 milhões de libras, pretende usar calor residual de data centers. Ele não estará funcionando antes de 2040 (ver quadro ao lado).

Dado o altíssimo custo do projeto e seu longo prazo, muitos observadores expressaram dúvidas à DCD se ele será realmente terminado.

"Londres definitivamente odeia ter ruas sendo escavadas", diz Kalita. "Imagine ter que escavar ruas para uma nova rede de calefação por toda a distância que ela precisará percorrer."

Outros tem uma questão ainda mais fundamental, e questionam se ela deveria mesmo ser feita.

"Não tenho familiaridade com o projeto para saber, mas quando vi o custo, olhei duas vezes", disse Charlie Beharrell da Heata. "O custo de conexão por residência, excluindo todos os custos de TI e custos excedentes, é mais do que o custo total de instalar um de nossos sistemas."

Os sistemas de que ele fala são vários, chamados de boilers digitais e caldeiras de dados.

A Beharrell, Junto com a Bjornsgaard e diversas outras empresas, estão apoiando uma abordagem diferente. Ela poderia ser mais eficaz do que as redes de calefação – mas ela repensa o data center.

A caldeira de dados

A ideia da caldeira de dados foi primeiro proposta pela Microsoft em um [artigo](#) de 2011, escrito por um grupo incluindo Sean James. A ideia básica é colocar racks de servidores nas casas e escritórios onde seu calor for útil.

O desenvolvimento da nuvem significa que trabalhos úteis podem ser enviados de forma criptografada para a caldeira de dados. Os clientes pagam um preço pela computação que cobre a eletricidade usada, e o host recebe calefação gratuita.

"Nos climas mais frios, aproximadamente 110 placas mãe poderiam manter uma residência tão aquecida quanto uma caldeira convencional faria", disseram os autores do artigo. "No resto do ano, os servidores ainda funcionariam, mas o calor gerado seria ventilado para o exterior, de forma tão inócua quanto uma secadora de roupas."

Várias empresas desenvolveram negócios considerando a ideia, incluindo a Garnot na França e a Cloud&Heat na Alemanha.

As empresas de calor distribuído de última geração e computação no Reino Unido inclui a Deep Green e a Heata.

A Deep Green coloca servidores de muita computação em tanques de imersão refrigerados a líquido de forma que sua energia calorífica possa ser canalizada para usuários próximos. Seus primeiros clientes são piscinas. A Heata oferece para residências eletrodomésticos que usam calor.

Ambas usam o mesmo provedor de nuvem, a Civo, para enviar trabalhos para suas unidades.

A ideia tem o potencial para desbancar as redes de calefação distritais, diz Beharrell, pois "é muito mais barato e fácil movimentar dados do que calor."

Caldeiras de dados podem ser colocadas em cada residência de um novo empreendimento imobiliário, por exemplo, com apenas uma necessidade mínima de infraestrutura. As residências já tem energia.

Algumas podem precisar de upgrades na rede, mas esse é um trabalho muito menor do que colocar tubulação a partir de um grande prédio de data center, através do sistema de calefação, para cada residência.

Beharrell não acredita que as duas abordagens sejam conflitantes, ele gostaria de ver projetos como a rede de calefação Park Royal irem para frente: "Acredito que é bom ser otimista e, honestamente, é bom para todo mundo se ele for para frente."

Mas ele gostaria de ver uma maior conscientização sobre a possibilidade de realmente distribuir a computação, ao invés de movimentar o calor: "Seria bom ver maiores investimentos em modelos mais disruptivos do que reforçar o status quo."

Decidir qual a melhor abordagem para uma dada situação requer um pensamento verdadeiramente estratégico.

"Alcançar as reduções de carbono necessárias até 2025 significa uma abordagem ao sistema completo", diz Wynn. "Integrar data centers nela seria uma grande contribuição para o sucesso." ●



Foto de Luke Stackpoole

CALEFAÇÃO DISTRITAL EM LONDRES

Este ano, o governo do Reino Unido deu um prêmio de 36 milhões de Libras (44,5 milhões de dólares) para um sistema de calefação distrital no Oeste de Londres, a ser construído pela Aecom, que compartilhará o calor residual de data center com até 10.000 novas residências.

Dois novos data centers em construção se comprometeram a fornecer 98,7 GWh de calor para o sistema, que não deve estar em operação antes de 2040.

O Fundo Green Heat Network (GHNF do Departamento de Energia pagará à Old Oak and Park Royal Development Corporation (OPDC) para criar o sistema de calefação para aproveitar calor residual a temperaturas entre 20°C (68°F) e 35°C (95°F).

O local é importante pois este é o local onde uma escassez na distribuição de energia elétrica efetivamente "paralisou" projetos de construção de residências em 2022, pois os data centers haviam reservado toda a energia disponível das subestações locais.

O projeto também receberá ao redor de 400.000 libras (495 mil dólares) de fundos para expertise técnica do programa Local Energy Accelerator (LEA) da Prefeitura de Londres, com David Lunts, CEO da OPDC, descrevendo-o como "um exemplo interessante e inovador do suporte da OPDC para as ambições de net zero da prefeitura." ●

Quando menos não é mais

Operadores de data centers precisarão adotar estratégias de refrigeração híbridas para lidar com as crescentes densidades de TI. O que isso significa?

'Menos é mais' é um destes clichês que raramente se mantém quando passam por um verdadeiro escrutínio, especialmente no que se refere a data centers. Os clientes sempre querem mais potência, mais espaço, mais de tudo, – ano após ano – e os operadores do data center precisam atender a essas demandas, sem tirar o olho da sua marcha rumo ao Net Zero.

Essas demandas são especialmente desafiadoras no que se refere à refrigeração do hall de servidores convencional, onde CPUs e GPUs mais potentes são amontoadas em racks, gerando mais calor e se traduzindo em uma charada para os projetistas maximizarem o fluxo de ar e o panorama térmico do data center.

De fato, com a demanda por aplicações de IA e de aprendizado de máquina devendo apenas se intensificar, racks com diversas GPUs se tornarão mais comuns para lidar com a maior carga de trabalho – mas o data center pode lidar com todo o calor que esse silício irá gerar?

Lutando por ar

"Todas essas novas demandas acrescentarão muita complexidade no panorama térmico", diz Nigel Gore, diretor sênior para refrigeração líquida na Vertiv.

Embora a refrigeração a ar tenha evoluído para lidar com os desafios sempre maiores apresentados a cada nova geração de tecnologia, ela enfrenta desafios para se adaptar para a próxima fase da computação.

A indústria resolveu a questão da refrigeração a ar para Alta Densidade movendo-a para mais perto da carga de calor nos racks, na forma de refrigeração na fila ou localizada no rack. O próximo passo foi gerenciar a recirculação do ar de

exaustão saindo para o corredor quente com temperaturas elevadas, para evitar o aumento da temperatura ambiente em todo o hall de servidores.

Entretanto, embora as tecnologias e as topologias de refrigeração a ar tenham sido engenhosamente adaptadas ao longo dos anos conforme cresciam as exigências de TI, os desafios são agora tão grandes que se chegou a um ponto de inflexão, onde as tecnologias de líquido precisarão ser implementadas para lidar no nível dos processadores com o aumento na potência.

Os operadores de data centers precisarão tomar uma decisão: como incorporarão a refrigeração líquida em seus halls de servidores para respaldar a TI de alta performance e, portanto, atrair e reter os clientes prósperos que dela precisam?

É fundamental que a instalação da tecnologia de refrigeração direta ao chip nas salas de servidores, de empresas como a Vertiv, seja simples. Os racks servidores ocupam o mesmo espaço do que os racks convencionais e podem funcionar junto com TI refrigerada de forma convencional, no mesmo hall.

Além disso, o maior 'risco' identificado que a refrigeração líquida apresenta – uma

parada resultante de um vazamento – foi reduzido significativamente com o design de rack com manifolds de líquido integrados que incorporam tubulação de alta qualidade. Essa tubulação inclui diversos engates, válvulas e seguranças intrínsecas para que os engenheiros possam realizar manutenções e upgrades nos racks, ou mesmo em servidores individuais localizados no rack, de forma rápida e segura.

De fato, se a sua impressão da refrigeração líquida foi tirada exclusivamente do mundo dos PCs, onde os entusiastas competem para construir o equipamento mais potente e com overclock que o seu dinheiro pode comprar (usando um cooler de água ou mesmo nitrogênio líquido em alguns casos extremos), então a realidade do rack servidor com refrigeração líquida moderna lhe surpreenderá.

Potência e água

Em uma instalação de rack normal, diz Gore, a primeira coisa a se verificar é a densidade de potência suportada. "Muitos data centers convencionais foram projetados ao redor de 5 kW por rack, mas agora estamos vendo 25, 50 kW e até já vimos 70 kW por rack", diz Gore. De fato, ele acrescenta, é possível que as densidades de potência logo cheguem a 100 ou 150 kW por rack já que a população de transistores deve crescer exponencialmente nos futuros chips de silício.

Abaixo do barramento de energia, normalmente é conectada a tubulação para refrigerante, de forma similar para cada rack. Para isso, a Vertiv usa tubulação de aço inox de alta qualidade para garantir que a corrosão galvânica não contamine o refrigerante e afete a transferência de calor nem a vazão. A empresa está também analisando desdobramentos em plásticos de alta temperatura como possíveis alternativas.

Essas redes de tubulação para refrigeração líquida, tecnicamente chamadas de rede secundária de fluidos (SFN), são projetadas com válvulas de corte para manutenção, com os acoplamentos dos racks de volta para a tubulação principal contendo válvulas in-line para isolar ainda mais o fluxo, se necessário. "Estes são acoplamentos manuais que podem ser desconectados, um design com uma válvula esfera que evita automaticamente qualquer gotejamento quando são desconectados", diz Gore.

Como alternativa, há válvulas eletrônicas que facilitam a instrumentalização do controle da vazão de refrigerante em conjunto com a unidade de distribuição de refrigerante, ele acrescenta.

Toda essa tubulação e hardware se conecta aos racks para fornecimento (refrigerante gelado) e retorno (refrigerante



Barramento de energia e manifolds suspensos para racks são componentes essenciais para dar suporte às implementações de refrigeração líquida

quente), e manifolds na fila possibilitam que, se necessário, racks individuais sejam isolados. Assim, o trabalho pode ser realizado em um rack – por exemplo, para instalar upgrades – enquanto os outros continuam funcionando.

Para halls de data center com piso elevado, essa rede de canos pode também ser instalada sob o piso. “Porém, pessoalmente eu acho que é mais fácil operar com a tubulação em cima se, por exemplo, você precisar desligá-la para fazer manutenção ou instalar upgrades no rack”, diz Gore.

Independentemente de onde você coloque a tubulação, é essencial que a pressão e a temperatura certas sejam mantidas. Para fazer isso, as unidades de distribuição de refrigerante – localizadas na fila, no rack ou no perímetro – usam monitoramento de vazão inteligente para controlar a vazão do líquido, bem como proporcionar a filtragem, portas facilmente acessíveis para enchimento de água e locais para drenagem.

“Também vimos implementações onde as equipes de facilities querem controlar automaticamente o líquido indo para cada rack, então, você pode ter controle de grupo e pode desligar a vazão para racks individuais, e pode não apenas desligar a vazão, mas também controlar o volume de vazão”, ele acrescenta.

Na parte traseira dos racks, cada servidor montado no rack tem acoplamentos para o refrigerante líquido – uma linha para fornecimento e retorno de água, codificadas por cores – enquanto os servidores têm um layout padronizado para as CPUs ou GPUs, fixando firmemente as placas de refrigerante diretamente sobre os chips.

Dentro do servidor, contatos de cobre proporcionam o melhor material condutor de calor para transferir de calor da CPU/GPU para o líquido, e enquanto a água retorna para a unidade de distribuição de refrigerante e o trocador de calor, ventiladores no servidor expõem o calor residual para o ar através da traseira do rack, no corredor quente do hall de servidores, para longe dos equipamentos de TI.

Refrigeração totalmente imersiva

Logicamente, a refrigeração líquida no nível do chip não é a única opção para

refrigeração líquida. A refrigeração por imersão permanece sendo uma opção, mas demandará uma curva de aprendizado para diversos operadores, com servidores imersos na horizontal – ao invés de na vertical – tanques em um líquido não condutivo cobrindo todos os componentes do rack e não apenas as CPUs e/ou GPUs. Vimos avanços nas especificações de líquidos dielétricos do Open Compute Project com a contribuição da Intel.

“Significa colocar todo o servidor em um líquido dielétrico com uma composição química específica. Como todos os componentes serão imersos – cada transistor, capacitor, módulo de memória, cabo, conector, etc. – é necessário ter certeza de que não há risco de contaminação, para garantir que seus servidores sejam compatíveis com o sistema”, diz Gore.

Por exemplo, cabos revestidos de PVC e plástico podem dissolver em líquidos ou óleos dielétricos e precisarão ser trocados por cabos ou plásticos com classificação para plenum. Material de referência é disponibilizado por fornecedores ou pelo Open Compute Project, fornecendo os requisitos para imersão e os guias de compatibilidade de materiais.

Também haverá perda de calor por radiação do tanque, que precisará ser suportada pela instalação. Diretrizes operacionais também avisarão sobre risco de contaminação sempre que um engenheiro abrir o tanque para fazer manutenção em um servidor ou para fazer um upgrade.

Além disso, quando comparando líquidos, nenhum deles pode chegar perto da eficácia da água no que se refere à física da transferência de calor.

“Ao examinar as propriedades de um fluido dielétrico comparado com a água por peso ou volume, há uma diferença importante em termos de suas propriedades e capacidade térmica. Com um fluido dielétrico, qualquer mudança na viscosidade afeta de maneira importante a capacidade de transportar calor. Como na refrigeração a ar, há também a questão complicada da “vazão de bypass”, onde obstáculos dentro do servidor evitam a vazão ótima nos componentes mais quentes, então, a refrigeração pode ser menos eficaz (do que a direta ao chip).

Então, há a necessidade de projetar a canalização do líquido para se ter certeza de que ele permanecerá um refrigerante eficaz”, diz Gore. Capturar otimamente o calor gerado enquanto o líquido é bombeado pelo tanque pode, portanto, ser um desafio em servidores e tanques densamente populados. Isso é bem compreendido, diz Gore, e as equipes de

desenvolvimento estão fazendo progressos nos dissipadores térmicos otimizados e no controle da distribuição da vazão interna.

Mas talvez onde a refrigeração direta ao chip seja muito melhor do que a refrigeração por imersão é simplesmente na utilização da água, ao invés de um óleo mineral não condutivo. “O líquido derradeiro antes de trocar para fluidos de refrigeração é essencialmente a água pois ela tem a maior capacidade térmica entre os líquidos. Com a refrigeração líquida, o que estamos fazendo é levar o líquido melhor condutor de calor diretamente aos componentes mais quentes no rack o mais rápido possível. A refrigeração por fluido está em nosso futuro como uma tecnologia emergente usando a mudança de fase para acelerar a remoção do calor”, diz Gore.

Com o aumento dos racks apoiando a refrigeração líquida, a adição da refrigeração líquida no data center já está ocorrendo. A Vertiv está desenvolvendo orientações e suporte para habilitar as instalações a acrescentar racks refrigerados a líquido, que podem coexistir com racks refrigerados a ar em halls de servidores, e podem aumentar conforme aumenta a demanda por hardware de servidores refrigerados a líquido.

À medida que a próxima geração de tecnologias dos chips chegue aos data centers, a questão da refrigeração líquida apenas se intensificará. Para colocar em perspectiva as características de geração de calor das CPUs e GPUs que estão na iminência de chegar, 750W são suficientes para alimentar um mini aquecedor com óleo que poderia manter uma pequena sala aquecida e deliciosa por todo o inverno.

Com oito GPUs de 750 W por servidor montado no rack, e entre 14 a 42 servidores por cluster, isso chega a quase 340 aquecedores por cluster. Multiplicados dentro um data center típico, os que não tiverem uma solução de refrigeração eficaz correm o risco de derreter.

>> Para saber mais sobre a tecnologia de refrigeração líquida da Vertiv, confira o [site da Vertiv dedicado às opções de refrigeração](#)



Linhas de fornecimento e retorno de refrigerante na refrigeração direta ao chip

Microfluídica: Refrigerando dentro dos chips



Peter Judge
Executive Editor

Se você pensa que os tanques de imersão são o destino final da refrigeração líquida, repense. A DCD escuta os engenheiros que querem refrigerante fluindo dentro de seus chips

Todos sabemos que a refrigeração líquida é o futuro para os data centers. O ar simplesmente não pode lidar com as densidades de potência que estão chegando nos data halls, então, fluidos densos com uma alta capacidade térmica estão entrando para assumir.

À medida que a densidade de calor dos equipamentos de TI aumenta, os líquidos se aproximam deles cada vez mais. Mas, quanto mais perto os líquidos podem chegar?

Ter um sistema de circulação de água nas portas traseiras dos gabinetes no data center tornou-se bem aceito. Depois, sistemas circulam água para placas frias em componentes especialmente quentes, como as GPUs e CPUs.

Além disso, sistemas de imersão submergiram racks inteiros em tanques de fluido dielétrico para que a refrigeração líquida pudesse entrar em contato com todas as partes do sistema. Os principais fornecedores agora oferecem servidores otimizados para a imersão.

Mas, há um passo à frente. E se o fluido pudesse ser levado para mais perto da fonte daquele calor – os próprios transistores dentro dos chips de silício? E se os refrigerantes fluíssem dentro dos processadores?

Husam Alissa, um diretor de tecnologia de sistemas na Microsoft, enxerga isso como uma opção futura interessante: “Em microfluídica, algumas vezes chamada de refrigeração embutida, 3D heterogênea, ou refrigeração integrada, levamos a refrigeração

para dentro do chip de silício, super perto dos núcleos ativos responsáveis pelo trabalho.”

Isso é mais do que apenas um sistema de refrigeração melhor, ele diz: “Quando você entra na microfluídica, você não está mais apenas resolvendo um problema térmico.” Chips com seu próprio sistema de refrigeração podem resolver o problema na fonte, no próprio hardware.

O nascimento da microfluídica

Em 1981, os pesquisadores David Tuckerman e R F Pease de Stanford sugeriram que o calor poderia ser removido de forma mais eficaz com minúsculos “microcanais” gravados em um dissipador de calor usando técnicas similares às usadas em fundições de silício.

Os pequenos canais têm uma área de superfície maior e removem o calor de forma mais eficaz.

O dissipador de calor poderia ser uma parte integrante dos chips VLSI, eles sugeriram, e a sua demonstração provou que um dissipador de calor de microcanais poderia suportar um então impressionante fluxo de calor de 800W por metro quadrado.

A partir daí, a ideia persistiu em universidades, mas apenas afetou tangencialmente a vida real dos chips de silício nos data centers.

Em 2002, os professores de Stanford Ken Goodson, Tom Kenny e Juan Santiago fundaram a Cooligy, uma startup com um design impressionante de “microcanais ativos” em um dissipador de calor incorporado

diretamente no chip, junto com uma bomba eletrocinética de estado sólido silenciosa e inteligente para circular a água.

As ideias da Cooligy foram absorvidas por partes das organizações tradicionais. A empresa foi comprada pela Emerson Network Power em 2005. A sua tecnologia, e parte de seus colaboradores, ainda circulam pela nova encarnação da Emerson, a Vertiv.

A ideia de integrar refrigeração e processamento se tornou mais prática à medida que a fabricação dos chips de silício se desenvolveu e passou para 3D. A partir dos anos 80, os fabricantes experimentaram construir múltiplos componentes um em cima do outro em um die de silício.

Fazer canais nos andares de cima de um chip de silício de múltiplas camadas é potencialmente uma vitória rápida para refrigerar, já que pode começar simplesmente pela implementação de minúsculos sulcos similares às aletas vistas nos dissipadores térmicos.

Mas a ideia não ganhou muita força já que os vendedores de chips de silício queriam usar técnicas de 3D para empilhar componentes ativos. Essa abordagem é agora aceita para memória de alta densidade e as patentes sugerem que a Nvidia pode estar pretendendo empilhar GPUs.

Na indústria de microprocessadores, a refrigeração e o processamento eram vistos como disciplinas separadas. Os chips precisavam ser projetados para dissipar seu calor, mas isso era feito de maneiras relativamente não sofisticadas, usando

materiais térmicos para sifonar calor para o grande dissipador térmico na superfície.

O dissipador térmico pode ser melhorado gravando-se canais menores, mas ele era um item separado, e o calor precisava atravessar uma barreira de adesivo para chegar lá.

Mas alguns pesquisadores podiam ver as possibilidades. Em 2020, Tiwei Wei, do Centro de Microeletrônica Interuniversidades e KU Leuven na Bélgica integrou a refrigeração e a eletrônica em um único chip.

Wei, cujo trabalho foi publicado na Nature em 2020, não achava que a ideia pudesse ‘colar’ para microprocessadores, dizendo que os microcanais de refrigeração seriam mais úteis em eletrônica de energia, onde chip de tamanho grande, feitos de semicondutores como o nitreto de gálio efetivamente administra e converte eletricidade dentro dos circuitos.

Essa possibilidade explica porque a Emerson / Vertiv queria adquirir a Cooligy, mas Way não via a tecnologia como indo além: “Esse tipo de solução de refrigeração embutida não é feita para processadores e chips modernos como a CPU”, ele disse à IEEE Spectrum.

Mergulhando nos chips

Naquele momento, os pesquisadores já estavam trabalhando na gravação de canais microfluídicos na superfície de chips de silício há alguns anos. Uma equipe na Georgia, trabalhando com a Intel em 2015, pode ter sido a primeira a fazer chips FPGA com uma camada de refrigeração microfluídica integrada, em cima do silício, “algumas centenas de microns (micrometros) distante de onde os transistores estão operando.”

“Eliminamos o dissipador térmico em cima do die de silício movendo a refrigeração líquida para apenas algumas centenas de micrometros de distância do transistor”, o chefe da equipe da Georgia Tech, Professor Muhannad Bakir, disse em comunicado de imprensa da Georgia Tech. “Acreditamos que integrar de forma confiável a refrigeração microfluídica diretamente no chip de silício será uma tecnologia disruptiva para uma nova geração de eletrônicos.”

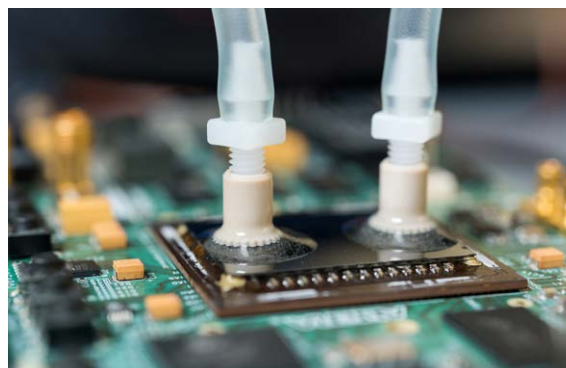
Em 2020, pesquisadores da École Polytechnique Fédérale de Lausanne, na Suíça, levou isso mais além, efetivamente passando fluido por túneis embaixo dos transistores geradores de calor.

O Professor Elison Matioli viu a oportunidade para colocar as coisas ainda mais próximas: “Projetamos a eletrônica e a refrigeração juntos desde o início”, ele disse em 2020, quando o trabalho de sua equipe foi publicado na Nature.

A equipe de Matioli conseguiu fazer a engenharia de uma rede tridimensional de canais microfluídicos de refrigeração dentro do chip propriamente dito, logo abaixo da

parte ativa de cada transistor, a apenas alguns micrometros de distância de onde o calor é produzido. Essa abordagem poderia melhorar a performance da refrigeração por um fator de 50, ele disse.

Matioli gravou fendas com largura de micrometros em uma camada de nitreto de gálio sobre um substrato de silício, e então, alargou as fendas no substrato de



silício para formar canais que fossem grandes o suficiente para ter um refrigerante líquido bombeado através deles.

Na sequência, as minúsculas aberturas na camada de nitreto de gálio foram vedadas com cobre e um dispositivo normal de silício foi criado no topo. “Apenas temos microcanais na minúscula região do wafer que fica em contato com cada transistor”, ele disse naquele momento. “Isso torna a técnica eficiente.”

Matioli conseguiu fazer dispositivos que usam muita energia, como um circuito retificador de 12 kV de CA para CC, que não precisam de dissipador térmico externo. Os microcanais levaram o fluido diretamente para os pontos de calor e trataram densidades de calor inacreditáveis de 1,7 kW por centímetro quadrado. Isso é 17 MW por metro quadrado, muitas vezes o fluxo de calor nas GPUs atuais.

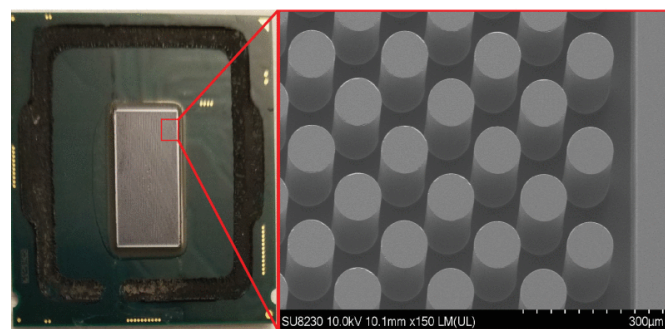
Rumo ao chip de silício padrão

Entretanto, o trabalho continua para adicionar microfluídica nos chips de silício padrão, através da criação de estruturas microfluídicas na traseira de microprocessadores existentes.

Em 2021, uma equipe liderada pela Microsoft, incluindo Husam Alissa, usou aletas de “micropinos” gravadas diretamente na parte de trás de uma CPU Intel Core i7-8700K de prateleira.

“Efetivamente pegamos um processador para desktop de prateleira e removemos a caixa”, diz ele. Sem a tampa que espalha calor e o material de interface térmica (TIM) o die de silício do chip ficou exposto.

“Quando esse die ficou exposto, aplicamos métodos de gravação para talhar os canais que queríamos ver”, continua ele. A parte



traseira do die foi gravada seletivamente, com uma profundidade de 200 microns, deixando um padrão de restolho nas hastes, com uma grossura de 100 microns – os “micropinos” que formam a base de um sistema de refrigeração direta ao chip.

Alissa avisa que esta é uma tarefa delicada: “Você precisa considerar com que profundidade está gravando, para não impactar as áreas ativas do chip de silício.”

Por fim, a traseira do die da CPU foi lacrada em um manifold

impresso em 3D, que entregava refrigerante para fluir entre os micropinos. O chip passava então por um overclock para dissipar 215 W de potência – mais do que o dobro da sua potência térmica de projeto (TDP), a energia a qual ele é projetado para suportar com segurança sem sobreaquecer.

Surpreendentemente, o chip foi capaz de funcionar nesse nível usando apenas água a temperatura ambiente. Entregue pelo manifold. O experimento mostrou uma redução de 44% na resistência térmica da derivação-para-entrada e usou um trigésimo do volume de refrigerante por watt do que teria sido necessário para uma placa fria convencional. A performance foi avaliada com programas de referência padrão.

Esta foi a primeira vez que canais microfluídicos foram criados diretamente em uma CPU padrão de consumo e alcançou a maior densidade de potência com a refrigeração microfluídica em um dispositivo CMOS ativo. Os resultados mostram o potencial para operar data centers de forma mais eficiente sem a necessidade de sistemas de refrigeração que usem muita energia, o grupo relatou na IEEE Xplore.

Tudo o que seria preciso é o fabricante de chips produzir em massa processadores com micropinos gravados e vendê-los empacotados com um manifold acoplado ao invés da tampa normal que espalha calor.

Se fundições como a TSMC pudessem fornecer seus chips com refrigeração líquida integrada, isto mudaria a dinâmica da adoção.

Também permitiria que a tecnologia ampliasse suas fronteiras, diz Alissa.

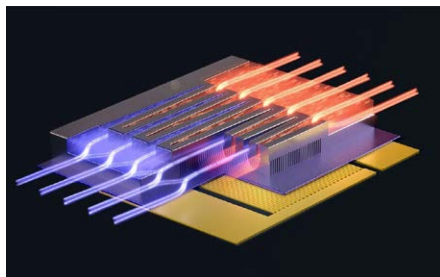
"Com as placas frias, você pode ter água a 40°C (104°F), mas com os microfluídicos você provavelmente poderia ter 80°C (176°F) ou temperaturas ainda mais altas saindo desses chips, pois o refrigerante está muito perto dos núcleos ativos", diz ele. "Isso logicamente aprimora a eficiência e os benefícios da recuperação de calor, em conjunto com menores requisitos para a taxa de vazão."

O futuro da microfluídica

"Há dois principais tipos de microfluídicos" diz Alissa. A opção amena, a qual ele diz poderá ser implementada em "alguns anos", é a abordagem que a sua equipe mostrou – gravar canais em chips comerciais: "Compre os chips, faça a gravação e pronto."

Uma versão mais completa dessa abordagem seria as fundições fazerem a gravação antes do chip chegar ao consumidor – pois nem todo mundo quer elevar a traseira de um processador e atacá-la com ácido.

Além disso, há o que Alissa chama da abordagem "pesada". Nela, você "intervém cedo nas fundições e começa a fazer estruturas em 3D." O que ele quer dizer são chips porosos que empilham componentes um em cima do outro com canais de refrigerantes nas camadas entre eles.



Há um projeto com base na abordagem usada por Matioli em Lausanne. Como diz Alissa, "Ela promete mais, mas, obviamente, é mais trabalhosa."

Alissa tem um objetivo: "O rumo que queremos tomar é o que nos permite otimizar esse chip ao mesmo tempo para a refrigeração e eletricamente, empilhando diversos dies um em cima do outro, com gravação (de microcanais) entre eles."

Se a refrigeração permite que vários componentes sejam empilhados em um die de silício fundo, conectados através de "vias pelo chip" (TCVs), conexões de cobre através do die de silício. Esses chips empilhados poderiam reduzir a energia e trabalhar muito mais rápido, já que os componentes estão mais próximos um do outro: "Em geral, ganha-se em performance, em refrigeração e também na latência por causa da proximidade."

Há outro benefício. Se os microfluídicos permitem que os chips passem para uma potência térmica de projeto (TDP) mais alta, isso poderia tirar uma das barreiras que os designers de chips de silício enfrentam.

A dificuldade de remover calor significa que os chips maiores atuais não podem usar todos seus transistores de uma vez só, ou sobreaquecerão. Os chips têm áreas de dark silicon (ver o quadro abaixo) e aplicar microfluídica poderia permitir aos designers deixá-los leves, impulsionando a performance do chip.

Mas não espere que a microfluídica resolva tudo. Em 2012, o Professor Nikos Hardavellas previu o próximo problema: "Mesmo se tecnologias exóticas de refrigeração fossem usadas, tal como refrigeração líquida acoplada à microfluídica, a entrega de energia para o chip provavelmente imporiria um novo entrave."

Uma vez que saibamos como remover mais calor do chip, teremos que desenvolver formas de entregar uma grande quantidade de energia que possa proporcionar integridade de sinal com as baixas tensões exigidas pelos transistores.

Estamos prontos para isso? ●

DARK SILICON

As gerações atuais e futuras dos chips têm um problema fundamental. A performance sempre melhorou à medida que mais transistores foram empacotados em um único processador. Mas agora, há tantos que eles não podem ser usados todos de uma vez, sem que o chip sobreaqueça.

Fabricantes de processadores publicam uma potência térmica de projeto (TDP) para cada chip, que é a quantidade de energia que ele pode suportar e dissipar com segurança – e presumiremos que há um bom dissipador de calor no chip.

As TDPs aumentaram muito. A GPU H100 SXM5 da Nvidia, por exemplo, tem uma TDP de 700W, que é enorme em comparação com CPUs padrão como as Xeons da Intel que consomem ao redor de 130 W.

Mas, o quanto você pode fazer com essa potência? Atualmente, os transistores fabricados com 4nm consomem minúsculos 10 attoJoules (10-18 joules) cada, então se um deles ligase em 1,8 GHz, consumiria 18 microwatts (18 x 10-9 W).

Isso é minúsculo, mas os atuais processadores tem uma quantidade colossal de transistores. Jon Summers do instituto de

pesquisas sueco RISE calcula que GPU H100 da Nvidia, que tem 80 bilhões de transistores, geraria 1.440 W - mais do que o dobro da TDP que a Nvidia publica para ela.

"Uma TDP de 700 W, deve significar que 51% do chip é dark silicon," Summers disse para o público do DCD Connect Londres em novembro de 2023.

A miniaturização continuada não consertará a situação. Transistores menores tem uma potência de ligação mais baixa, então, mais pode ser aceso dentro da faixa da TDP, mas a quantidade de transistores também está aumentando.

Summers diz que a Intel planeja ter um trilhão de transistores em um chip até 2030, cada um usando ao redor de 1 aJ por ligação. Se a velocidade do clock aumentou para 4 GHz, o chip é 1.000 mm quadrados, e fluxo térmico, então isso significa que 40% dos transistores precisam ficar 'escuros'.

Agora, as TDPs são baseadas em um fluxo máximo de calor que pode ser removido de um chip. A H1000 da Nvidia tem uma área de 814 mm quadrados, Então o fluxo de calor é de 860 kW por metro quadrado.

Isso é comparável aos níveis vistos em demonstrações de fusão nuclear e Summers espera que a Intel pressione para 2,4 MW por m².

A questão do dark silicon é conhecida por muito tempo: Em 2012, o Professor Nikos Hardavellas da Northwestern University, disse em uma revista da Associação de Computação Avançada, Usenix: "Sem um milagre tecnológico, caminhamos para uma era de "dark silicon", capazes de construirmos dispositivos densos que não podemos energizar. Sem a capacidade de usar mais transistores ou fazê-los mais rápidos, melhorias de performance devem estagnar a não ser que mudemos de direção."

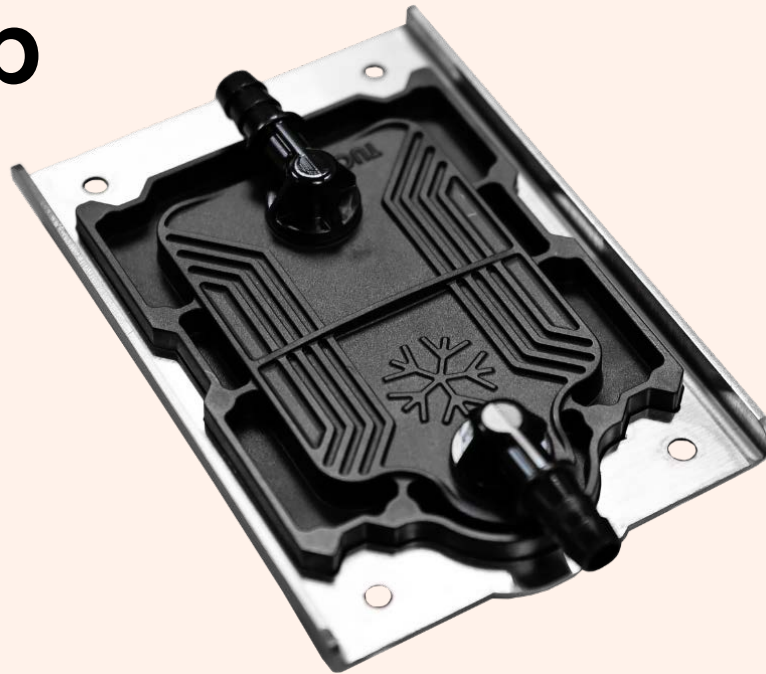
Tem havido inúmeras abordagens para o problema, sobretudo aumentando o uso de núcleos especializados dentro dos chips, que são usados apenas quando necessário.

Mas talvez uma forma de reduzir o dark silicon seja se os fluidos pudessem fluir dentro do chip propriamente dito, onde eles pudessem remover mais energia e permitir mais fluxo de calor. ●

Refrigeração de precisão no nível do chip



Sebastian Moss
Editor-in-Chief



O CEO da JetCool sobre como milhares de minúsculos jatos podem estar chegando em um data center perto de você

A medida que as temperaturas dos chips e as densidades dos racks aumentam, uma infinidade de empresas vem a público apresentar a sua visão do futuro.

As necessidades de refrigeração da inteligência artificial e outras cargas de trabalho de alta densidade ultrapassaram a capacidade dos sistemas a ar, demandando algum tipo de refrigeração líquida.

"Quanto pensamos no panorama da refrigeração líquida, vemos três categorias técnicas diferentes," explicou o CEO da JetCool, Bernie Malouin.

"Há a imersão de uma única fase, mergulhando no óleo. E ela é interessante, mas há algumas limitações na potência do chip – por muito tempo, eles ficaram parados em 400 W. Alguns estão tentando um pouco mais, mas não tanto quanto é necessário."

"Vemos esse tipo lidando com os processadores maiores (potência térmica

de design (TDP)), e esses podem chegar a 900-1000 kW. Eles são tecnologicamente adequados para o futuro da computação, mas são freados pelos componentes químicos."

Muitas soluções de duas fases usam substâncias polifluoroalquil (PFAS), conhecidas como 'químicos eternos' que são relacionados a riscos para a saúde humana e enfrentam restrições nos EUA e na Europa. Empresas como a ZutaCore prometeram trocar para outra solução até 2026, mas a mudança tem sido lenta.

"É uma preocupação para vários de nossos clientes, eles estão nos procurando pois estão preocupados sobre a segurança daqueles fluidos", disse Malouin.

"Eles estão preocupados com a disponibilidade continuada desses fluidos. "E há a terceira categoria: as placas frias de Refrigeração Líquida Direta (DLC). "Nós somos dessas", disse Malouin. "Há outros".

Placas frias de DLC são uma das formas mais antigas de refrigeração líquida para TI – simplesmente movimentando líquido gelado para



placas metálicas montadas diretamente nos componentes mais quentes. Elas têm sido usadas há bastante tempo pela comunidade de computação de alta performance, mas a JetCool acredita que está na hora de renovar o conceito.

Ao invés de passar fluido sobre a superfície, seus jatos de resfriamento direcionam o fluido diretamente para a superfície de um chip. “Temos esses conjuntos de mil minúsculos jatos de fluido e trabalhamos diretamente com os principais fabricantes de chips – as Intels, AMDs, NVidias – e arrumamos esses jatos de forma inteligente para que se alinhem para o local onde as fontes de calor estão em um dado processador.”

Ao invés de tratar o chip como um todo com uma necessidade singular de refrigeração, a abordagem de refrigeração microconectiva “tenta balancear as cargas de calor disparatadas, requisitos térmicos diferentes de determinadas partes daquele empilhamento de chips”, disse Malouin.

“Quanto você começa a pensar sobre pacotes verdadeiramente integrados, os núcleos talvez possam funcionar a uma temperatura um pouco mais alta, mas então, você pode ter seções de memória com maior largura de banda (HBM) que não precisam de tanta energia, mas têm um limite menor de temperatura.”

Ao invés de tentar fazer o design para núcleos de alta potência e os HBMs sensíveis a temperatura, cada seção pode ser refrigerada a uma taxa diferente. “Isso permite que você separe essas coisas e tenha refrigeração de precisão onde precisa”, disse Malouin.

Embora Malouin acredite que a refrigeração líquida no nível da instalação é o futuro dos data centers, a empresa tem também um sistema independente para aqueles buscando mergulhar o pé em águas mais frias, com uma parceria com a Dell focada em implantações de duas cavidades.

Dois módulos de bombeamento pequenos proporcionam a circulação de fluido e um trocador de calor a ar ejeta calor no outro lado do sistema Smart Plate.

“Quando adicionamos essas bombas, acrescentamos algum consumo elétrico, mas não precisa que os ventiladores trabalhem com tanta força, então fica de 15 a 20 decibéis mais quieto – e extraímos ao redor de 100 W por servidor após retirarmos a desvantagem das bombas”, afirmou Malouin.

Quando você trabalha com 10 racks ou mais, ir para o nível da instalação faz mais sentido, ele disse. Perguntado sobre a temperatura de entrada preferida, Malouin disse que o sistema era flexível, mas acrescentou: “nós na verdade realmente gostamos de fluidos mornos.”

Ele disse: “Temos instalações hoje que estão alimentando temperaturas de entrada de refrigeração de 60°C (140°F) e mais altas. E ainda estamos refrigerando esses dispositivos sob carga plena.” Isso ainda não é comum, mas Malouin acredita que águas mais quentes serão mais populares em lugares como a Europa devido ao potencial de reutilização do calor.

Nos EUA, a empresa é parte do projeto COOLERCHIPS do Departamento de Energia, que busca

avanços substanciais nos sistemas de refrigeração de data centers.

O foco do prêmio de mais de 1 milhão de dólares da JetCool é não apenas sobre o potencial da refrigeração, mas tem um fascinante benefício secundário: “Temos casos onde fizemos o chip de silício intrinsecamente entre 8 e 10% mais eletricamente eficiente”, afirmou Malouin.

“Isso não tem nenhuma relação com o uso de energia pelo sistema de refrigeração, mas com vazamentos”

Malouin não quer dizer vazamentos do sistema de refrigeração, e sim o fenômeno quântico das correntes de vazamentos dos semicondutores que podem ter um impacto significativo na performance de um chip.

A história recente da refrigeração de data centers tendeu a presumir que permitir que as temperaturas aumentassem mais economizaria energia porque menos seria usada para refrigeração. Resultados, incluindo a pesquisa de Jon Summers no instituto de pesquisa sueco RISE, estão indicando que as correntes de vazamento no silício limitam os benefícios do funcionamento mais quente.

“Uma grande parte do nosso empreendimento COOLERCHIPS é para substanciar isto através de evidências científicas mais rigorosas e extrapolar para diferentes ambientes para ver onde se aplica e onde não.”

Olhando ainda mais adiante, Malouin vê uma oportunidade para ir mais fundo no chip de silício. “Em alguns casos, ele pode ser efetivamente integrado como uma camada embutida dentro do chip de silício, e então acoplado a um sistema que seja externo e que esteja fazendo alguma reutilização de calor. Quando pensamos sobre o assunto de uma forma holística, pensamos que há uma oportunidade real para um salto qualitativo na eficiência do data center.”

Por ora, a empresa diz que é capaz dar suporte a cargas de 900 W das GPUs maiores da Nvidia e está atualmente refrigerando chips ‘feitos sob medida’ não revelados que usam 1500 W.

“Por fim, você realmente precisará dar uma olhada na refrigeração líquida se quiser rodar não apenas o futuro da IA generativa, mas se quiser rodar o agora da IA generativa.” ●

Por que a Intel está desenvolvendo tecnologias de refrigeração para data centers



Sebastian Moss
Editor-in-Chief

Processadores de 2kW e o futuro dos data centers

Costumava ser tão fácil.. A Intel não é uma empresa de refrigeração, e não quer ser uma. Mas a empresa está sendo

A empresa de chips se tornou cada vez mais preocupada que os chips do futuro sobrecarregarão os data centers do futuro e está pesquisando agora como pode ajudar.

“Queremos estar prontos: Como a potência dos processadores continua a crescer, os desafios de refrigeração crescem exponencialmente”, explicou Tejas Shah, engenheiro chefe do grupo de plataformas de supercomputação. Em resposta, a Intel está tentando melhorar a refrigeração do data center.

Crescentes densidades de potência têm levado os desenvolvedores a usar refrigeração por imersão e alguns estão olhando para a imersão em duas fases onde fluidos em ebulição removem ainda mais calor. Shah lidera uma iniciativa relativa a esses sistemas com resistência térmica ultra baixa, dissipadores térmicos com formato de coral integrados com uma cavidade de câmara de vapores em 3D.

O projeto foi um dos 15 escolhidos pela COOLERCHIPS, um programa para mudar drasticamente como os data centers são refrigerados, gerido pela ARPA-E (Agência de Projetos de Pesquisa Avançados – Energia) dos EUA.

Mas antes de entrarmos na tecnologia, vamos abordar o elefante na sala: Fluidos de imersão de duas fases provaram ser controversos: “Não estamos desenvolvendo o fluido, usaremos fluidos padrão da indústria”, disse Shah.

“Estamos muito conscientes do problema com as PFAS e as regulamentações que as circundam”, ele admitiu, fazendo referência aos tóxicos ‘químicos eternos’ usados pela maioria das soluções de duas fases.

“A 3M está praticamente saindo do negócio de imersão em duas fases, estamos trabalhando com alguns dos fornecedores de fluidos.”

A outra questão é o potencial de aquecimento global (GWP) de muitos dos fluidos, que pode superar quaisquer economias com eficiência energética que a imersão pudesse trazer. “Tirar as PFAS é o maior desafio, mas estamos confiantes de que podemos atender aos requisitos de GWP”, disse Shah.

Agora, voltamos a tecnologia. A empresa está construindo dissipadores térmicos com formato de coral usando manufatura aditiva, ou seja, impressão em 3D.

“É basicamente otimizar o design para seguir uma função matemática específica”, disse Shah. “Então, a otimização de sua topologia – é um procedimento matemático iterativo para otimizar para uma determinada função de custo, qualquer que ela seja. Nesse caso, está reduzindo a resistência térmica e garantindo que acertemos os pontos de calor certos.”

Os dissipadores térmicos são acoplados com cavidades de câmara de vapores em 3D: bolsos de metal planos e vedados preenchidos com fluido, para espalhar a capacidade de ebulição do sistema de duas fases.

“O formato de coral está aumentando a área da superfície, mas também queremos reduzir a resistência de espalhamento, que é o motivo pelo qual temos uma câmara de vapores embutida dentro da estrutura em 3D”, disse Shah.

Quando a corrente flui através de material semicondutor, ela encontra “resistência de espalhamento”, onde os elétrons se espalham por uma área mais larga do que inicialmente previsto. Baixar a temperatura de um chip ajuda a reduzir o espalhamento e, portanto, a resistência.

A nova abordagem também usa um revestimento aprimorado, o qual melhora a ebulição ao promover a nucleação. “Então, ela ajuda bastante a ebulição ocorrendo na superfície”, disse Shah.

Com as três abordagens combinadas, a empresa pode resfriar processadores com até uma potência térmica de projeto de 2 kW. Isso é além do que está no mercado hoje, ou provável de ser lançado nos próximos anos, apesar do Grace Hopper Superchip da Nvidia ter uma TDP de 450 W a 1 kW combinando a CPU e a GPU.

O sistema de imersão pode dar suporte a “pelo menos oito processadores”, consumindo 2 kW cada, disse Shah. Aumentar a quantidade em um rack, passa a ter uma limitação de energia ao invés de uma limitação térmica, ele disse.

Embora os testes tenham se mostrado promissores, superar cada problema não é uma conclusão inevitável – especialmente se o setor de imersão não puder resolver o problema das PFAS (o fornecedor ZutaCore diz que eliminara as PFAS até 2026, mas ainda precisa fornecer os detalhes específicos, enquanto a rival LiquidCool está diversificando para imersão de uma única fase).

“Não é uma vitória garantida, é um grande desafio”, disse Shah a respeito do projeto e de outras iniciativas da Intel. “Mas a meta é reduzir a resistência térmica e atingir a próxima geração de processadores com a melhor solução térmica que exista.”

Entretanto, se funcionar, a Intel pode se encontrar em uma posição interessante de ser um fornecedor de refrigeração. “O que poderíamos acabar fazendo – é um grande poderíamos – mas poderíamos inicialmente até vender a solução térmica com o processador”, disse Shah.

“Isso construirá o ecossistema e, então, a próxima geração poderá se equiparar aos nossos processadores.”



A IA é quente.

Garanta que o seu data center fique frio.

Adote hoje as estratégias de refrigeração híbrida para lidar com as crescentes densidades de TI.



Veja como podemos ajudá-lo em **[Vertiv.com/GetCool](https://www.vertiv.com/GetCool)**