# Desafios do Processamento de Alto Desempenho

**Philippe Olivier Alexandre Navaux, Matheus da Silva Serpa**

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

{navaux, msserpa}@inf.ufrgs.br

***Abstract.*** *Driven by the need for high processing capabilities required by Big Data, Machine Learning and Artificial Intelligence algorithms, High-Performance Computing has rapidly become one of the most active computer science fields. The state-of-art algorithms from these fields are notoriously demanding in terms of computer resources. Choosing the right computer system to optimize their performance is paramount. This paper presents the main challenges on future computer architectures, heterogeneous systems and quantum computing.*

***Resumo.*** *O Processamento de Alto Desempenho, High Performance Computing - HPC, vem crescendo de importância nos últimos anos com a necessidade de grande poder de processamento para gerenciar sistemas Big Data, assim como para treinamento e uso de algoritmos para Inteligência Artificial, entre outras demandas. Neste artigo é feita uma análise dos principais desafios na evolução destes sistemas para os próximos anos, quanto a futuras arquiteturas dos processadores e máquinas, a heterogeneidade, a computação quântica, além das necessidades de redução do consumo de energia e localidade dos dados e o avanço de soluções HPC na nuvem.*

## 1. Introdução

Processamento de Alto Desempenho, do inglês *High Performance Computing* (HPC), tem ao longo dos anos sido a área de especialistas que se preocupam com as máquinas que possuem o maior poder de processamento de determinada época, representada pelos supercomputadores. Tradicionalmente ocorrem anualmente duas seleções que indicam quais máquinas no mundo alcançaram o topo deste poder de processamento que é divulgado no TOP500 [J. Dongarra and Strohmaier 2021].

Este cenário está mudando nos últimos anos, pressionado pelas necessidades de maior poder de processamento dos algoritmos de Inteligência Artificial, *Machine Learning*, *Deep Learning*, assim como da necessidade deste poder para melhor gerenciar os dados nos ambientes de Big Data [Verbraeken et al. 2020]. Em resumo, o HPC não é mais uma área que atende um nicho de necessidades de pesquisadores nas áreas da física, química e biologia, entre outras específicas, mas sim uma área que precisa atender demandas da sociedade como um todo.

Nesse sentido, este artigo apresenta estas demandas, assim como quais são os desafios que a área de HPC enfrenta e tenta resolver. Serão apontadas a evolução das arquiteturas de máquinas cada vez mais heterogêneas, as demandas de energia, a necessidade de os dados serem cada vez mais locais, a computação na nuvem entre outros tópicos necessários para a evolução da área [Serpa et al. 2019].

## 2. Demandas da área de Big Data

A quantidade de dados gerados por ano dobra a cada dois anos e cresce exponencialmente [Desjardins 2019]. Estamos chegando ao patamar do Yotta dados $10^{24}$ ou $1.000^{8}$. Por outro lado, apenas da ordem de 20% desta informação seria útil. Muitos destes dados têm origem em sensores, em IoT (*Internet of things*). Tal quantidade de dados, conhecida por *Big Data*, necessita de tratamento por sistemas não tradicionais para manipulação, análise e extração de informações deste conjunto de dados.

Em 1997, foi empregado pela primeira vez o termo *Big Data,* referindo-se ao crescente número de dados gerados a cada segundo no mundo de forma estruturada ou não estruturada. Michael Cox e David Ellsworth, ambos trabalhando na NASA, escreveram o artigo "*Managing Big Data for Scientific Visualization*" para a *Conference on Visualization* de 1997, apresentando o conceito de *Big Data* à comunidade acadêmica [Cox and Ellsworth 1997].

Portanto, Big Data é uma coleção de conjuntos de dados tão grande e complexa que se torna difícil de processar usando ferramentas de gerenciamento de banco de dados usuais. Muitas vezes é uma coleção de dados legados. A área Big Data, além de tratar de novas formas de gerenciar dados, precisa de um grande poder de computação, para conseguir processar estes dados. Além disso, devemos lembrar que dados são números, códigos sem nenhum tratamento. Já a informação são os dados tratados. É o processamento dos dados que vão criar um significado. Por fim, o conhecimento é o saber sobre determinado assunto, é ter uma aplicação para a informação.

Também é importante colocar, que nos dias de hoje o conhecimento sobre a informação é poder, sempre foi, mas com o advento da grande quantidade de dados e a capacidade de processar estes, tornou-se possível fazer análises e tomadas de decisões. A informação tornou-se o bem mais importante. Empresas detentoras de informações e poder de processá-las e analisá-las detém hoje um Poder muitas vezes maior do que os Estados. Além do que, não existem fronteiras para a informação e estas empresas estão captando informações em quase todo o mundo.

Concluindo, máquinas com grande capacidade de processamento tal como os supercomputadores, são necessárias para auxiliar no armazenamento e extração de dados na era do *Big Data*.

## 3. Demandas da Área de Inteligência Artificial

Segundo o relatório "*AI for Science*" do *Department of Energy* dos USA [Stevens et al. 2019] novas técnicas de Inteligência Artificial (IA) serão indispensáveis para suportar o crescimento contínuo e expansão da infraestrutura da Ciência através de sistemas *exascale*. A experiência da comunidade científica, o emprego de *Machine Learning*, a simulação com HPC, os métodos de análise dos dados, permitiram um crescimento único e novo de oportunidades para Ciência, novas descobertas e mais poderosos métodos aceleraram a ciência e suas aplicações em benefício da humanidade.

A convergência de HPC com a Inteligência Artificial permite que ambientes de simulação empreguem o aprendizado por reforço profundo em diversos problemas como simulação de robôs, aeronaves, veículos autônomos, etc. Técnicas de *Deep Learning*

(DL), aprendizado profundo, estão acelerando as simulações pela substituição dos modelos em áreas como previsão de clima, geociência, fármacos, etc. Novas fronteiras em física estão sendo alcançadas pelo aumento da aplicação de Equações Diferenciais Parciais (EDP), com o emprego de DL para simulações.

Por outro lado, um dos grandes gargalos no emprego de algoritmos para *Machine Learning*, *Deep Learning*, é a necessidade do treinamento destes algoritmos, o aprendizado, antes do seu emprego. Esta atividade pode levar semanas, dependendo da infraestrutura computacional disponível. É onde o processamento de alto desempenho acelera esta etapa, viabilizando o emprego destes algoritmos.

## 4. Desafios do Processamento de Alto Desempenho

As duas últimas seções, trataram sobre Big Data e Inteligência Artificial, mostrando o quanto as necessidades de avanços nestas duas áreas estão demandando cada vez mais poder de processamento para executar os modelos, para gerenciar os arquivos, para extrair os dados, entre outras demandas.

Nas seções abaixo, serão apresentados os principais desafios que o HPC precisa vencer para obter resultados em tempo hábil para seu emprego. Um exemplo disso é a etapa de treinamento de algoritmos de *Machine Learning*, acima apontada, que pode consumir um tempo muito grande se não forem executados em máquinas mais rápidas.

### 4.1. Nova geração de Processadores e Aceleradores

Recentemente, diversas propostas de novas arquiteturas paralelas surgiram, o que deve agitar e alterar o mercado de processadores. A empresa ARM, que está sendo objeto de aquisição pela NVIDIA, projetou o processador A64FX que permitiu ao computador FUGAKU da *Fujitsu* [Fujitsu 2021], Japão, alcançar o primeiro lugar na última versão do TOP500, com desempenho de 442 Petaflops, podendo chegar ao *Exaflop* na versão atualizada da máquina (Figura 1).

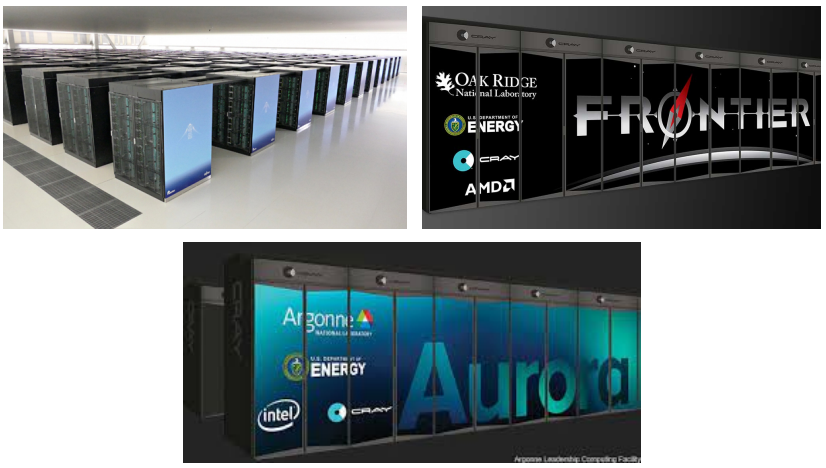


**Figura 1 - Supercomputador Fugaku, hoje no TOP500 e futuros Frontier e Aurora.**

Já nos USA, está previsto o lançamento para fins de 2021 de um supercomputador chamado de *Frontier* [Frontier 2021], no laboratório de Oak Ridge,

fabricado pela Cray-HPE, que deve alcançar 1,5 *Exaflops* e que usará os processadores *AMD Epyc* e aceleradores *Radeon*. No laboratório de Argonne, está previsto para 2022, a instalação de uma máquina, de nome *Aurora* [Aurora 2021], que chegará também ao *Exaflop*, empregando processadores *Intel* adotando a arquitetura *Ponte Vecchio* que integra processador *Xeon* e acelerador *Xe* no mesmo *chip*. Verifica-se que cada vez mais as arquiteturas destes processadores e máquinas são heterogêneas permitindo que a execução das aplicações seja processada na arquitetura com o melhor desempenho.

### 4.2. Arquiteturas Heterogêneas

Os novos processadores, que já estão surgindo no mercado, possuem cada vez mais arquiteturas heterogêneas, isto é, são compostos por CPUs e GPUs em um mesmo *chip* [Davila et al. 2019], tais como os da Intel (Ponte Vecchio) e os da AMD (APUs). Outra possibilidade são o SoC, System on Chip, onde no mesmo *chip* (Figura 2) estão processadores, acelerador, memória e sistema de E/S. Em geral estes chips são empregados em ambientes junto a sensores, IoT, para *edge computing*. Além desta opção no *chip,* existe a possibilidade de a heterogeneidade estar no nível da placa, onde convive o processador com a GPU, ou então com a FPGA (Intel A10).

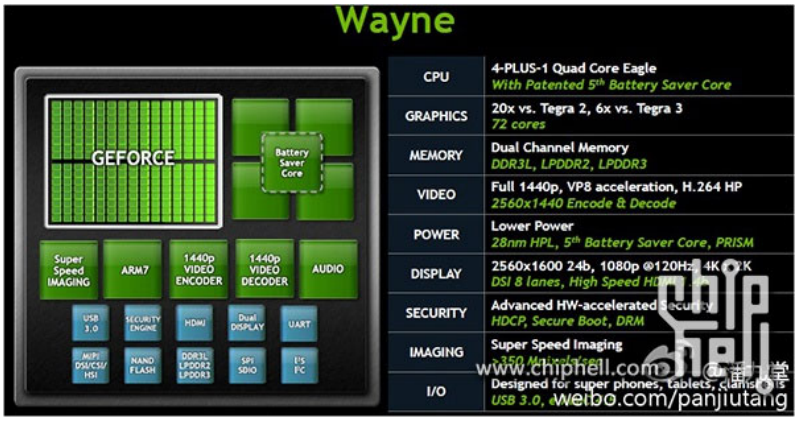


**Figura 2 - Chip Tegra4 da NVIDIA integrando processadores com GPUs num SoC.**

Neste ambiente cada vez mais heterogêneo, com diferentes aceleradores, o ambiente de programação deverá mudar de um atendimento de um determinado tipo de acelerador para estar preparado para um ambiente de programação para atender diferentes tipos de aceleradores [Vetter et al. 2018]. Por exemplo, com o surgimento das GPUs da AMD (Big Navi) e da Intel (Xe-HPG) a programação destas deverá ser feita numa linguagem universal diferente de CUDA (proprietária da NVIDIA) tal como o OpenCL.

### 4.3. Bibliotecas de Programação Paralela

Com o surgimento das diversas arquiteturas de computadores para memória compartilhada, distribuída e computação heterogênea, diferentes linguagens de programação e bibliotecas para processamento paralelo foram introduzidas na literatura. Na Figura 3, extraída de [Diaz et. al 2012], é possível ver que no ano de 2003, com o

surgimento dos processadores multi-core, a biblioteca mais utilizada era a MPI, para memória distribuída. Entretanto, após alguns anos, o uso da biblioteca OpenMP para memória compartilhada foi aumentando. Em 2008, com a revolução das placas gráficas do tipo GPU, a biblioteca CUDA teve um aumento no número de citações, junto com a biblioteca OpenCL, também utilizada para programação em sistemas heterogêneos.

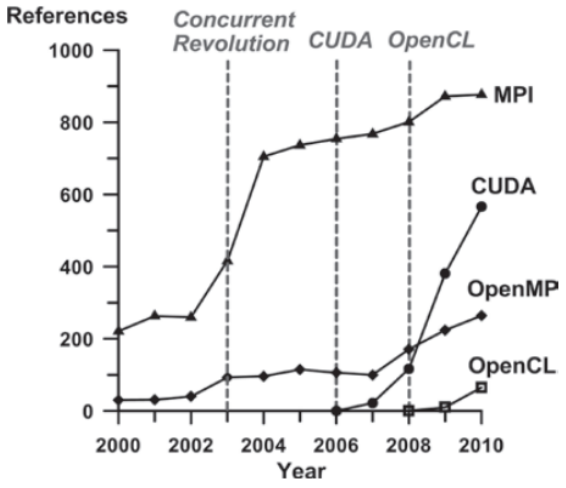


**Figura 3 - Quantidade de citações das principais bibliotecas de programação paralela ao longo dos anos [Diaz et al. 2012].**

Uma das formas de explorar paralelismo em processadores multi-core e many-core é através do uso da memória compartilhada. Nesse tipo de programação, os núcleos computacionais acessam a memória compartilhada diretamente, comunicando-se através dela. Geralmente, o modelo *fork-join*, é utilizado, no qual várias *threads* são criadas em determinado momento, chamado de *fork*, e em outro momento, chamado de *join*, todas *threads, exceto* a inicial, deixam de existir.

A biblioteca POSIX Threads, por ser uma implementação leve de *threads*, foi uma das primeiras a ser utilizada para criação de aplicações paralelas [Barney 2009]. Entretanto, devido a complexidade e a necessidade do programador de decidir diversos aspectos do sistema operacional, a mesma deixou de ser usada para esse fim e continuou sendo usada apenas para o desenvolvimento de aplicações voltadas ao sistema operacional. Em substituição à biblioteca POSIX Threads, surgiu a biblioteca OpenMP, que utiliza diretivas do tipo pragma para definir o paralelismo de maneira simples e rápida [Chandra et al. 2001]. Outras bibliotecas como a Cilk Plus [Schardl et al. 2018], desenvolvida pelo MIT e após, mantida pela Intel, surgiram para facilitar o uso da programação vetorial e de tarefas, entretanto, devido a evolução da biblioteca OpenMP, também deixaram de ser utilizadas.

Com a popularidade dos sistemas heterogêneos formados por processadores e placas gráficas do tipo GPU, bibliotecas como CUDA [Cook 2012], OpenCL [Munshi et al. 2001] e OpenACC [Farber 2016] surgiram. A biblioteca CUDA consiste de um conjunto de extensões para C, C++ e FORTRAN, que permitem criar *kernels*, que são funções que podem ser executadas em placas gráficas do tipo GPU da NVIDIA. Além dessa biblioteca, a OpenCL também surgiu como uma *framework* para computação

heterogênea, o qual permite desenvolver aplicações que executam tanto em processadores multicore quanto em placas gráficas de qualquer fornecedor. Por fim, temos a biblioteca OpenACC, que tem se tornado muito popular, por ser parecida com a biblioteca OpenMP. A mesma também utiliza *pragmas* e torna a programação para GPUs simples e rápida.

Quando supercomputadores com milhares de nós computacionais são utilizados, uma biblioteca para comunicação distribuída é necessária. A biblioteca MPI, proposta em 1992, surgiu com o intuito de utilizar a troca de mensagens para realizar a comunicação entre múltiplos nós computacionais [Gabriel et al. 2004].

Aplicações paralelas atuais e eficientes tendem a utilizar tanto a biblioteca MPI para troca de mensagem entre os diferentes nós computacionais, quanto as bibliotecas OpenMP e CUDA / OpenACC, para uso dos múltiplos núcleos de um processador e das placas gráficas do tipo GPU. A tendência é o surgimento e evolução dos compiladores, tanto no sentido da auto paralelização quanto na otimização da localidade dos dados.

### 4.4. Computação Quântica

Com os processadores quânticos começando a se tornarem realidade, já é possível imaginar máquinas heterogêneas que possuírão unidades de processamento quântica. A perspectiva de conseguir operadores do tamanho de um átomo, com capacidade de operar usando técnicas de transistores de germânio [Hendrickx et al. 2021], está viabilizando a chegada ao mercado dos processadores *Qubits*. Com isto as futuras arquiteturas teriam unidades de processadores X86 com unidades *qubits* (Figura 4).

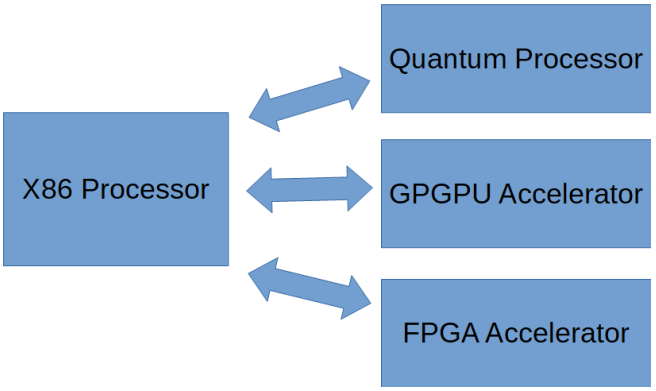


**Figura 4 - Futuro das Arquiteturas Heterogêneas com Processadores Quânticos.**

Nesta perspectiva, os processadores quânticos serão empregados como aceleradores, num primeiro momento. Estes processadores resolverão problemas especialmente em áreas de segurança, criptografia, meteorologia, fármacos, biotecnologia, modelos econômicos entre outros.

A Computação Quântica será a próxima fronteira nas mudanças da capacidade de processamento, esta capacidade permitirá resolver problemas da Ciência, hoje

inviáveis de solução em tempo hábil.

### 4.5. Necessidades de Energia

A crescente demanda de poder de processamento pelas máquinas HPC, levou os fabricantes a associarem milhares de processadores em clusters, que geram um consumo de energia elevado. Algumas máquinas do TOP500 chegam a consumir na casa dos 30 MW, correspondendo ao consumo de energia elétrica de uma cidade de cerca de 300.000 habitantes, isto levou aos construtores de máquinas, assim como os fabricantes de processadores a otimizarem as arquiteturas para que o consumo diminua. Estas otimizações passam pela alteração da arquitetura do processador, gerenciar as unidades não ativas, diminuir o relógio do processador, entre outras técnicas [Padoin et al. 2019]. Atualmente as máquinas não procuram somente o aumento na velocidade na execução das instruções, mas também uma diminuição no consumo de energia, o que às vezes entra em oposição.

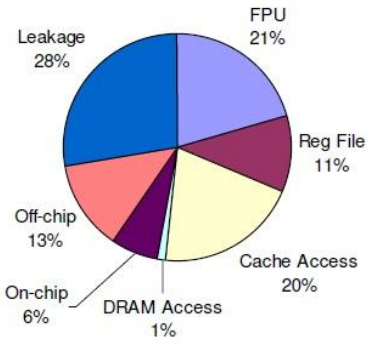


**Figura 5 - Distribuição do consumo de energia num *core* [Kooge et al. 2008].**

A Figura 5, acima, de um relatório para o DARPA [Kooge et al. 2008], demonstra claramente um dos desafios que existe para a evolução das arquiteturas dos processadores. Verifica-se que o percentual da energia dedicada para o efetivo processamento, execução da instrução, é cerca de 20% de toda energia, enquanto que cerca de 28% é gasto na energia estática, *leakage*, que é a energia gasta pelo circuito estar ligado, mesmo sem nada executando. Outros 20% é gasto pelo acesso a cache e 11% nos Registradores. Mesmo que estes percentuais tenham melhorado, verifica-se que a energia investida no objetivo final, que é a execução da instrução, é pequena comparado com a energia gasta nas outras partes do funcionamento do processador, este é um desafio importante a ser melhorado nas futuras arquiteturas dos processadores.

### 4.6. Localidade dos Dados

Nos futuros microprocessadores a energia empregada para a movimentação dos dados terá um efeito crítico no desempenho destes. Qualquer nano-joule de energia empregado para mover dados para cima e para baixo na hierarquia da memória irá diminuir a energia disponível para a computação. Mapeamento de tarefas e escalonamento precisam ser otimizados na rede de interconexão, priorizando a localização [Cruz et al. 2021]. Isto implica em que a movimentação de dados deve ficar

restrita ao máximo possível. Portanto, a tendência é priorizar a localidade dos dados em detrimento da velocidade do processador, embora dados locais tendem a auxiliar no processamento mais rápido. A conservação da energia torna-se a prioridade.

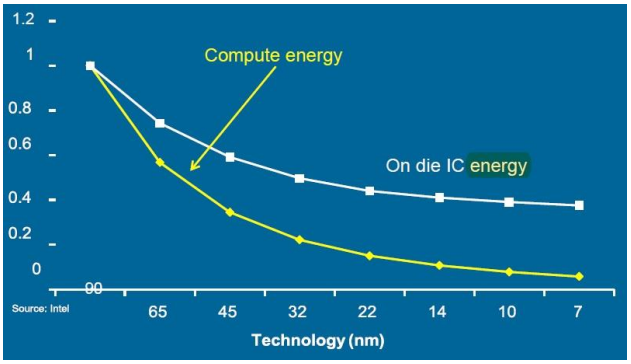


**Figura 6 - Consumo de energia em Pico Joules versus evolução da tecnologia [Vetter et al. 2018].**

A Figura 6 extraída do relatório do DoE [Vetter et al. 2018] mostra que apesar da evolução da tecnologia dos chips, com tecnologias cada vez mais fina chegando ao 7nm, a diminuição do consumo de energia de todo chip não acompanha a diminuição da energia devido à computação. Isto mostra que o consumo de energia para a movimentação dos dados consome mais energia que a execução das operações de computação no chip. Isto implica numa revolução importante, não só na arquitetura dos processadores, mas também na forma de executar as instruções. Os compiladores deverão preocupar-se em deixar os dados mais próximos aos processadores.

## 4.7. Resiliência

Resiliência lida com a habilidade de um sistema em continuar operando na presença de falhas ou de flutuações de desempenho. Supercomputadores, que processam aplicações de alto desempenho, possuem milhares de cores, memórias e circuitos os conectando e portanto a probabilidade de ocorrer uma falha em algum dos seus elementos torna-se uma certeza com a quantidade de circuitos.

Uma análise ao longo do tempo mostra o crescimento das falhas com a evolução dos supercomputadores, com cada vez mais cores, chegando ao milhar destes. A Figura 7 mostra esta evolução, onde as linhas mostram a variação considerando diferentes taxas de falhas [Kooge 2008]. Verifica-se que com uma taxa de falhas de 0,01 no ano a probabilidade de ter uma interrupção é de algumas horas. Isto quer dizer que um supercomputador com milhares de processadores vai ter servidores parando por falhas todos os dias, donde a importância da resiliência para manter a máquina funcionando.

A Resiliência é obtida através de hardware e software que irão dinamicamente detectar a falha, diagnosticar, reconfigurar e reparar esta, permitindo o processamento continuar sem percepção pelo usuário. Esta área torna-se essencial nas máquinas futuras para atender as necessidades do processamento de alto desempenho e permitir que este processamento possa acontecer até obter os resultados sem interrupções.

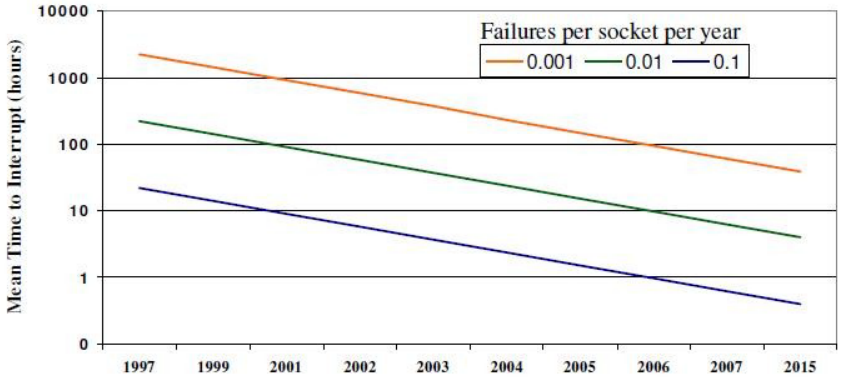


**Figura 7 - Evolução das falhas por circuitos ao longo dos anos [Kooge et al. 2008].**

## 4.8. Computação na Nuvem

Com o aumento da demanda de HPC para diversas áreas, como acima mencionado Big Data e Inteligência Artificial, os grandes provedores de computação na nuvem começaram a interessar-se em prover estas facilidades. Observou-se o surgimento de instâncias com processadores mais poderosos, com GPUs, com FPGAs e uma melhoria do sistema de interconexão dentro da nuvem para permitirem aos usuários instanciar um conjunto de máquinas com condições de atender uma demanda de maior poder de processamento (Figura 8).

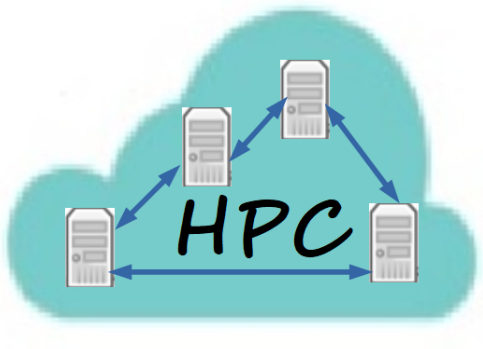


**Figura 8. Cloud HPC.**

Nos sites dos principais provedores de cloud encontram-se anúncios tais como "*High Performance Computing on AWS Redefines What is Possible*", "*Cray in Azure - a dedicated supercomputer on your virtual network*", "*Build your high-performance computing solution on IBM Cloud*", "*Google Cloud - HPC in the cloud becomes reality*" mostrando claramente a importância e interesse que estas empresas estão dando a oferecer HPC na *cloud*. Projeções indicam que daqui a uns 2 anos este mercado deve superar os 30 bilhões de U$, representando parte importante e crescente de HPC.

## 5. Conclusão

Como abordado no texto acima, o Processamento de Alto Desempenho passou de uma área específica, atendendo determinadas necessidades de processamento, para um tema

central na evolução da computação, considerando as necessidades crescentes de poder de processamento de áreas como Big Data e Inteligência Artificial entre outras.

Esta evolução passa por transformações importantes das máquinas e processadores, inclusive com o emprego crescente da nuvem, cloud, para atender estas demandas de poder de processamento. A busca pelo maior desempenho, nem sempre é a principal prioridade, hoje, muitas vezes, procura-se otimizar o consumo de energia. A heterogeneidade é parte integrante dos processadores e das máquinas e a chegada dos processadores quânticos deve aumentar esta diversidade. A resiliência e as novas formas de programação e armazenamento são parte essencial deste desenvolvimento. Conclui-se que a evolução da computação passa pelo crescimento contínuo do poder de processamento, mas também por novas formas em fazê-lo.

## Agradecimentos

Este trabalho foi financiado pelos projetos Petrobras 2016/00133-9 e GREEN-CLOUD: Computação em Cloud com Computação Sustentável (#16/2551-0000 488-9), da FAPERGS e do CNPq, programa PRONEX 12/2014. Além disso, os autores gostariam de agradecer o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

## Referências

Aurora (2021). Argonne Leadership Computing Facility. https://www.alcf.anl.gov/aurora [Acesso em: 10 Abr. 2021].

Barney, B. (2009). POSIX threads programming. National Laboratory. https://computing.llnl.gov/tutorials/pthreads [Acesso em: 4 Mai. 2021].

Chandra, R., Dagum, L., Kohr, D., Menon, R., Maydan, D., & McDonald, J. (2001). Parallel programming in OpenMP. Morgan kaufmann.

Cook, S. (2012). CUDA programming: a developer's guide to parallel computing with GPUs. Newnes.

Cox M., Ellsworth D. (1997) “Managing Big Data for Scientific Visualization” Conference on Visualization ’97. VIS ’97.

Cruz E., Diener M., Pilla L., Navaux P. (2021) “Online Thread and Data Mapping Using a Sharing-Aware Memory Management Unit” ACM Transactions on Modeling and Performance Evaluation of Computing Systems January 2021, Vol. 5 No. 4 Article.

Davila G. P. , Oliveira D. A. G. , Navaux P. O, A, , Rech P. : Identifying the Most Reliable Collaborative Workload Distribution in Heterogeneous Devices. DATE 2019: 1325-1330

Desjardins J. (2019). How much data is generated each day? https://www.visualcapitalist.com/how-much-data-is-generated-each-day/. [Acesso em: 12 Mar. 2021].

Diaz, J., Munoz-Caro, C., & Nino, A. (2012). A survey of parallel programming models and tools in the multi and many-core era. IEEE Transactions on parallel and distributed systems, 23(8), 1369-1386.

Dongarra J., H. M. and Strohmaier, E. (2020). Top500 supercomputer: November 2020. https://www.top500.org/lists/top500/2020/11/. [Acesso em: 10 Mar. 2021].

Farber, R. (2016). Parallel programming with OpenACC. Newnes.

Frontier (2021). ORNL Exascale Supercomputer. https://www.olcf.ornl.gov/frontier/ [Acesso em: 10 Abr. 2021].

Fujitsu (2021). Supercomputer Fugaku. https://www.fujitsu.com/. [Acesso em: 10 Abr. 2021].

Gabriel, E., Fagg, G. E., Bosilca, G., Angskun, T., Dongarra, J. J., Squyres, J. M., ... & Woodall, T. S. (2004, September). Open MPI: Goals, concept, and design of a next generation MPI implementation. In European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting (pp. 97-104). Springer, Berlin, Heidelberg.

Hendrickx N., Lawrie W., Russ M., Riggelen F., Snoo S., Schouten R., Sammak A., Scappucci G., Veldhorst M. (2021) "A four-qubit germanium quantum processor". Nature, 2021; 591 (7851): 580 DOI: 10.1038/s41586-021-03332-6.

Kooge P. & All (2008) "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems" Defense Advanced Research Projects Agency Information Processing Techniques Office, Tech. Rep.

Munshi, A., Gaster, B., Mattson, T. G., & Ginsburg, D. (2011). OpenCL programming guide. Pearson Education.

Padoin E. L., Diener M., Navaux P. O. A. , Méhaut J. F. : Managing Power Demand and Load Imbalance to Save Energy on Systems with Heterogeneous CPU Speeds. SBAC-PAD 2019: 72-79

Schardl, T. B., Lee, I. T. A., & Leiserson, C. E. (2018, July). Brief announcement: Open cilk. In Proceedings of the 30th on Symposium on Parallelism in Algorithms and Architectures (pp. 351-353).

Serpa, M. S., Moreira, F. B., Navaux, P. O., Cruz, E. H., Diener, M., Griebler, D., & Fernandes, L. G. (2019). Memory Performance and Bottlenecks in Multicore and GPU Architectures. In 2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP) (pp. 233-236). IEEE.

Stevens R., Nichols J., Yelick K., Helland B., Leads C. (2019) "AI for Science" Report on the Department of Energy (DOE).

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., & Rellermeyer, J. S. (2020). A survey on distributed machine learning. ACM Computing Surveys (CSUR), 53(2), 1-33.

Vetter J. and all (2018) "Extreme Heterogeneity 2018: Productive Computational Science in the Era of Extreme Heterogeneity" Report for DOE ASCR Basic Research Needs Workshop on Extreme Heterogeneity.