Artigo: Por que usar o Linux na Bioinformática?

Clovis Ferreira dos Reis
PhD Bioinformática

Autor: Clovis F. Reis
E-mail: cfreis230@gmail.com

Artigo: Por que usar o Linux na Bioinformática?

Objetivo: Apresentar algumas razões que levam a escolha do Linux como o principal Sistema Operacional para a Bioinformática.

Sumário

Por que usar o Linux na Bioinformática?

1 Introdução

"Esqueça o Windows e instale o Linux". Esta é, provavelmente, uma das primeiras frases ouvidas em um curso de bioinformática. Costuma chocar e causar indignação em muita gente, já que ter que abandonar o Windows e migrar para o Linux soa-lhes como um verdadeiro absurdo. Rechaçam veemente a ideia e, dispostos a provar o seu ponto de vista, instalam, compilam e adaptam as ferramentas necessárias ao seu trabalho, exibindo, orgulhosamente, a plena possibilidade de utilização do seu amado Sistema Operacional (SO).

Mas claro, logo começam a surgir alguns dissabores e "dores de cabeça" onde os processos demorados e pesados típicos da bioinformática insistem em transformar o computador em uma vernissagem de Yves Klein [1], com uma sequência interminável de telas azuis.

Como *"ultima ratio regum"* instalam o Linux em uma máquina virtual e como efeito colateral convertem tanto o SO hóspede quanto o hospedeiro em carroças inúteis.

Por fim, com os ânimos abalados e prazos atrasados, acabam por se render e, humildemente, aceitam remover o Windows, instalar uma das diversas distribuições Linux e reestabelecer a paz com os deuses da Bioinformática.

2 Confiabilidade, estabilidade e segurança

Pois bem, após esta polêmica introdução e antes que seja tarde, quero deixar claro que este artigo não tem como objetivo causar controvérsia acerca de qual é o melhor SO, quais suas vantagens e desvantagens. Visa sim constatar o fato de que bioinformática de qualidade e Linux andam de mãos dadas e apontar algumas das razões que tornam isto uma realidade.

Comecemos com um fato muito revelador: consultando os dados disponibilizados pelo site Top500 [2] que semestralmente realiza o levantamento das configurações dos Top 500 supercomputadores existentes no mundo, verificamos que, das 500 máquinas mais avançadas do planeta, 500 rodam sobre o Linux. Ou seja 100% utilizam o Linux e 0% utilizam outros SO.

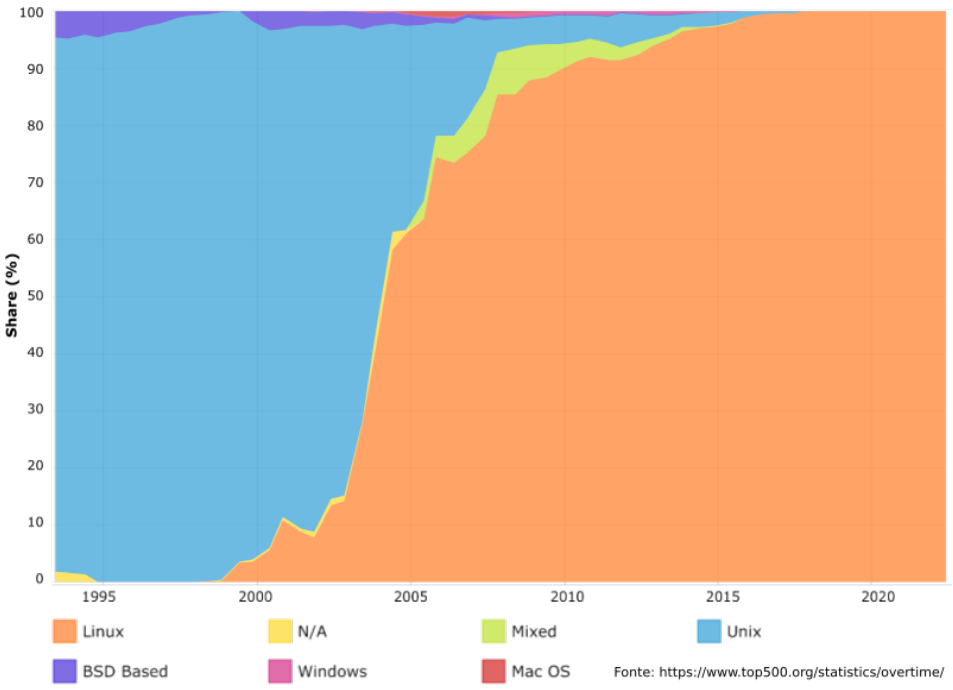


Figura 1: Utilização de sistemas operacionais nos Top 500 supercomputadores do mundo no decorrer do tempo

Realizando um levantamento histórico, com dados disponibilizados a partir de 1993, podemos constatar que esta hegemonia ocorre desde junho de 2004, tendo o Linux atingido a marca de 100% em novembro de 2017, como pode ser visto na Figura 1. Já o Windows nunca teve uma marca superior a 0,8%. A evidente conclusão a que se chega é que onde há computação de alto desempenho, há Linux.

Coincidência? Claro que não. Mas o que isso significa?

Significa que a realização de processamento massivo, de alto desempenho e confiabilidade necessita de uma plataforma robusta, estável e segura, que hoje só pode ser proporcionada pelo Linux. E se é bom para a computação de alto desempenho, é ótimo para a bioinformática. Simples assim. Mas porque o Linux é a escolha óbvia de tantos especialistas?

3 Justificativas

3.1 Código aberto

Comecemos pelas justificativas mais clichês no universo dos SO. Em primeiro lugar na lista, e em último no quesito originalidade, podemos afirmar que o Linux é um SO de código aberto e, como tal, possui um enorme suporte de toda a sua comunidade.

Em segundo lugar destacaríamos a questão pouco inédita, mas neste caso bastante importante do custo. Se considerarmos que um supercomputador é uma máquina enorme, composta, muitas vezes de centenas ou milhares de máquinas menores, devemos também lembrar que cada um destes chamados nós estará rodando uma instância do SO, passível de licenciamento. Por exemplo, o Sunway TaihuLight, supercomputador chinês hoje considerado o terceiro mais rápido do mundo, possui uma rede contendo algo em torno de 650.000 CPUs e 10,5 milhões de *cores* [3]. Nestes casos, os custos das licenças de uso poderiam tornar-se um grande entrave administrativo, cujo inconveniente é prematuramente eliminado pela simples utilização do Linux, já que ele é totalmente isento de royalties.

3.2 Modularidade e estabilidade

Mas estes são pontos menores do nosso argumento, havendo outras questões ainda mais relevantes, não só para os supercomputadores, mas também para qualquer máquina para a bioinformática. O código aberto e facilidade na realização de ajustes em parâmetros internos do SO tornam o Linux uma plataforma altamente personalizável, permitindo a sua adaptação à qualquer necessidade específica. É também um SO de alta escalabilidade, permitindo que uma mesma máquina se adapte a diversos regimes e cargas de trabalho. Porém, a meu ver, a superioridade do Linux sobre o Windows desponta de forma indubitável quando consideramos a sua modularidade e estabilidade.

Para que não haja confusão de conceitos, aqui entende-se por estabilidade a capacidade que o SO tem de permanecer funcionando por um longo período de tempo, sem qualquer necessidade de intervenção, após ser instalado e corretamente configurado. Quando lidando com servidores isto é, na minha opinião, fundamental. Após instalado, o sistema vai sendo lenta e gradualmente ajustado às necessidades do serviço prestado, num trabalho que pode levar meses até que seja obtida aquela sintonia fina perfeita. Assim, um SO onde existe a real necessidade de reinstalações ou recuperações periódicas nos impedirá de alcançar um estado de produtividade plena das demais ferramentas que rodam sobre a plataforma.

E no quesito estabilidade o Linux prova-se campeão. A maior complexidade e grau de conhecimento necessários à sua instalação e configuração são plenamente compensados pelos meses consecutivos de operação sem sustos e ausência de *reboots*, recuperações ou reinstalações recorrentes. Do outro lado temos um SO cujo fabricante afirma oficialmente em seu website: “se você já tentou praticamente tudo que podia, e o computador ainda não está funcionando, a restauração talvez possa corrigir o problema”.

E como *pièce de résistance* vem a modularidade do Linux, que o torna a plataforma ideal para nós, bioinformatas. Quem algum dia já trabalhou com bioinformática sabe da colossal capacidade de processamento necessária à execução de algumas tarefas, onde cada bit de memória e cada FLOP de processamento é importante. Como consequência, quanto menor for o custo computacional necessário para manutenção do SO, melhor. Apesar de ambos, Windows e Linux possuírem kernels monolíticos e modulares, a capacidade deste último no quesito modularidade é insuperável.

A Figura 2 apresenta um esquema resumido do grau de modularização de cada um dos SO. Em vermelho encontram-se assinaladas as funções do SO que não podem ser “desligadas” e em azul os módulos que, optativamente, podem ser ativados. No Linux praticamente tudo pode ser desativado, dependendo da sua necessidade, com uma economia de recursos considerável. Na maioria dos casos, sendo a interface gráfica completamente dispensável, simplesmente não se inicia o *display server* e economiza-se uma infinidade de recursos.

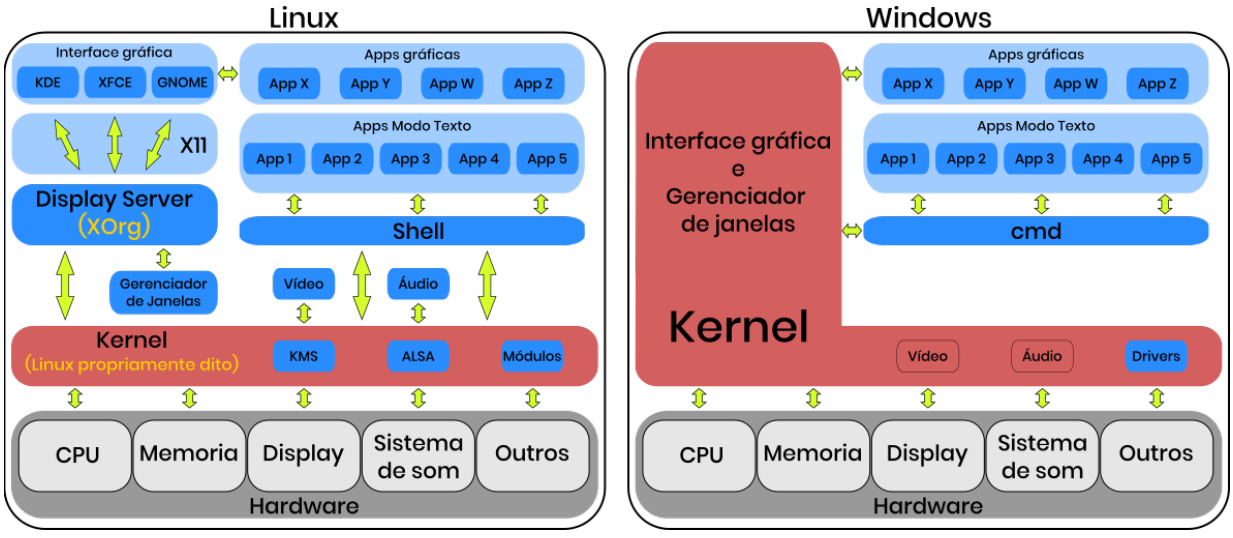


Figura 2: Comparativo entre a modularidade do Linux e do Windows

Mas se fizermos questão de utilizá-la, como quando rodamos algum processo de bioinformática em segundo plano em nossa própria estação de trabalho, existem muitas alternativas econômicas disponíveis, já que no Linux até as interfaces gráficas são intercambiáveis. Assim, é só escolher uma interface extremamente leve dentre as diversas disponíveis, e o consumo de recursos basais pela máquina ainda será baixo. Sistemas de som, modos

complexos de exibição em vídeo e outras perfumarias também podem ser desativadas, concentrando os recursos da máquina nas tarefas que realmente interessam. Em casos extremos, versões personalizadas e enxutas de kernel podem ser compiladas, sendo até mesmo o shell passível de supressão, mas muito raramente é preciso chegar-se a tanto.

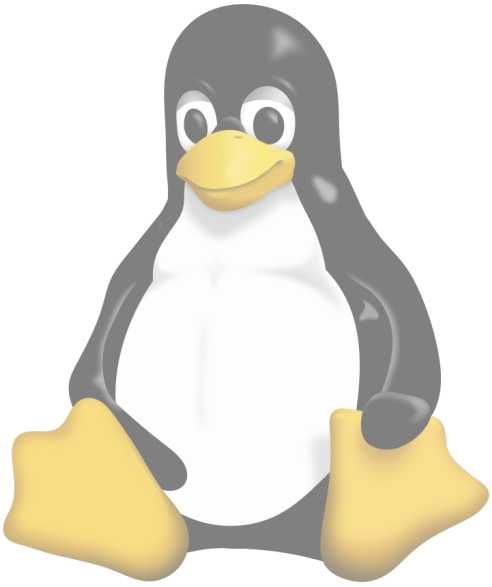
Já o Windows foi pensado de forma a simplificar as tarefas administrativas necessárias, voltado-se para um público que precisa saber muito pouco para começar a operá-lo. Desta forma, os módulos opcionais estão concentrados, de uma maneira geral, sob a forma de drivers de dispositivos que podem ser carregados sob demanda, muitas vezes sem a anuência do usuário. A interface gráfica não é opcional e irá consumir boa parte dos recursos da máquina. Até mesmo sua interface de comando está subordinada a esta interface gráfica, sendo dela dependente. Dezenas de processos serão inciados em segundo plano, muitas vezes à revelia do administrador do sistema, não havendo uma opção segura para desativá-los sem causar a instabilidade de todo o SO. Em resumo, a vida vegetativa do Windows acaba por consumir recursos preciosos que poderiam estar sendo utilizados em nossa tarefa fim e não na manutenção da infraestrutura computacional.

3.3 Linhas de comando

Como um bônus, o Linux ainda oferece o shell script, uma ferramenta fantástica e poderosa capaz de resolver muitos problemas de bioinformática de forma simples e eficiente. Tarefas como concatenar e filtrar aquivos imensos são corriqueiras em nosso dia-a-dia. Se no Windows a manipulação personalizada de arquivos vai depender do desenvolvimento de uma ferramenta específica desenvolvida em Python ou outra linguagem qualquer, no Linux o problema pode muitas vezes ser resolvido com alguns cats, greps, seds e pipes bem encadeados, com uma economia de tempo considerável.

4 Conclusão

Desta forma, podemos afirmar com absoluta certeza que o Linux é a escolha óbvia quando se pensa em plataforma para bioinformática. Oferecendo uma solução robusta, estável, modular e customizável, assegura ao profissional da área a liberdade e tranquilidade para se concentrar na sua atividade fim, sem perder a noite de sono preocupado se o SO irá travar nos quinze minutos derradeiros daquelas vinte sete horas gastas em um alinhamento. Claro que você, caro leitor, pode discordar da minha opinião, mas tenha em mente que 500 dentre os 500 administradores dos Top500 supercomputadores do mundo concordam comigo.



Referências

[1] Wikipédia. Yves klein. https://pt.wikipedia.org/wiki/Yves_Klein. [Online; acessado 02/09/2022].

[2] TOP 500. Development over time. https://www.top500.org/statistics/overtime/. [Online; acessado 02/09/2022].

[3] Statista. Number of computer cores in the 10 fastest supercomputers in the world in 2022. https://www.statista.com/statistics/268280/number-of-computer-cores-in-selected-supercomputers-worldwide/. [Online; acessado 02/09/2022].