LINUX + GENOMICS

TWO REVOLUTIONS

Martin Krzywinski

USENIX 2004 Conference

I wrote this narrative to collect my thoughts for the talk. You can find the slides at

http://mkweb.bcgsc.ca/talks/linuxgenomics

TITLE

Without doubt, Linux and Genomics have both made a difference. I don't need to tell you how, where and what kind of a difference Linux has made. The fact that we're here makes the point. What I will tell you is how genomics has made a difference and how it is likely to impact your life. For the better.

ACKNOWLEGEMENTS

I work with about 150 people who use Linux on a regular basis. For programming. For checking email. For command-line acrobatics. Many of the people on this list live at the prompt. Others work in the lab, carry out experiments and make sure that lights on a lot of very sexy machinery continue to blink.

At this point I'd like to extend very special thanks to Yaron Butterfield. Both his and my name appear on this talk, and only I am here. We were going to tag-team this talk. Yaron unfortunately cannot be here today.

KIRK

Just like Linux and genomics, Star Trek (original episodes) was filled with forward-thinking, optimistic ideas. The ideas that the future in hold will be better. That one day, in the same room, different races and species will work towards making life better, in the ways that life should be made better. Pursuit of knowledge, equality, and empowerement. I'm sure that in the bowels of the Enterprise, there is a very large UNIX computer.

REVOLUTION

What's a revolution? I think we all have our definitions. Some of you may be thinking banners. Some of you public protests and pamphlets printed late into the night in some dark basement. I'd like to argue that both Linux and Genomics are two revolutions. Revolutions that happened because a better way could be found and because sufficient technology existed to support that way. And, of course, that happened because a handful of people had a vision.

LINUX

The Linux revolution gave power to the computer user. It provides a better way. Progress slows down in a monopoly. Many of us face this everyday. Linux opened the tap on progress. Change. Something new and different, at times half-baked but already much tastier than other stale choices.

REVOLUTION 1

I can dust off an old 386, install a distribution of Linux and provide network services for hundreds, if not thousands, of people. Linux offers me a context to start, contribute and play with computer science projects. Learn about programming, systems, databases, graphics, networking. If I'm stuck I can Google my problems away.

I don't need to persuade you that Linux will succeed because it rewards innovation and sharing – fundamentally positive and self-sustaining behaviours. Linux has already succeeded.

REVOLUTION 2

The genomics revolution is providing biology with its fundamental underpinning. Like quantum mechanics of the early 1900's, biology has discovered in the last 10 years its own framework. Like Linux, genomics grew up with the computers, Perl and the web. LAMP is genomics' right arm – Linux, Apache, MySQL and Perl/PHP. Scientists in this field started by sharing data and ideas by email, ftp and web. They did it because that's the way it's always been done. At least in a young field like genomics which really started in about 1990.

In 15 years the human genome has been sequenced. The mouse, rat, and chicken genomes are available, along with many others. Significant progress has been made in understanding how genes operate, interact and cause or provide resistance for disease. And a lot of this work has been achieved on Linux.

COMMON PHILOSOPHY

While I'm trying to relate Linux and Genomics, it's interesting to note that science and programming are similar activities. Both activities have components which can be directly related. Although I may be a little ambitious in saying that programming improves the human condition, I think it does. People do it because it's fun, because it makes them happy. The same can be said about science. I do science because it makes me happy. I write Perl scripts because it makes me happy. Maybe one day something I do will make a difference. Of this I'm sure: if enough people do what makes them happy, eventually someone will achieve something extraordinary.

COMMON PROCESS

Programming and Genomics share a process. The development of code and the steady progress of science both require robustness, logic and elegance. Simplicity in models and power in their predictive powers. Short, fast algorithms that solve the problem or get you closer to a solution.

OPENNESS

Linux and Genomics are similar in that they embrace openness. Now Linux and Genomics do not hold the monopoly on this. But, I'm talking about Linux because this is a Linux sig and I'm talking about genomics because that's what I do. Perhaps if I were at a Windows conference and I were a virus creator, I'd be making a similar point.

Science thrives on open source principles. It requires them. Just like contributing to a repository of code, scientists contribute to the sphere of knowledge. Sometimes the sphere grows, its edges are never well defined. Sometimes it flutters a little bit when a paradigm shift occurs, like quantum mechanics, which makes us look at the same thing in a different way.

Science needs openness. There is no humanity in science – science is a hypothesis-driven process. But we do science and we are human. What we uncover will impact our lives, this planet. Science needs to be open so that meaningful dialogue about its discoveries can take place. Even with so much information, so many channels for its dissemination, it can be argued that the average individual knows relatively less about science today than 50 years ago. Perhaps science is moving too fast. Perhaps it's too esoteric. Whichever – we need to focus on science HOWTOs, and READMEs.

PUBLIC EFFORT

When science is open, science moves forward faster than when science is closed. A great example of this is the assembly of the human genomic sequence. In June of 2000, there was a close race to assemble the sequence – between the HGP and Celera. Jim Kent from UCSC used a cluster of Linux boxes to process and combine the sequence reads and the world had its first human genome assembly. And it was public. You could download it and print it out.

ACTGs

I'm talking about Linux and Genomics and most of you know more about Linux than I do. This slide fails to cover all of genomics very well. Genomics is the large-scale study of the content and role of DNA. Most of us have 46 chromosomes. About 2-3% of these chromosomes have special regions called "genes". These regions are used by the biochemistry of the cell to make proteins. The proteins participate in biochemical reactions and do very important things. There wouldn't be life as we know it without proteins.

Sometimes the DNA on a chromosome changes. It may change due to age, exposure to radiation, a chemical or just bad luck. Most of the time the damage is repaired automatically. Sometimes it is not. In this case, a single 3-base pair codon is missing from a part of chromosome 7. If you are missing these three base pairs you have cystic fibrosis.

HISTORY

Revolutions start quickly and flourish quickly. Linux started in 1991 and has rapidly progressed through a large number of production versions. Genomics started at pretty much the same time and it has progessed through equally staggering changes. Sequencing capacity has grown by 100x fold over the last 10 years. We went from sequencing tiny bacteria to huge mamallian genomes.

During this time computers appeared. First in our offices, then in our homes, then in our pockets. Later all these computers would be connected and we could have a movie streamed to the computer in our pocket.

GROWTH RATE

The moral of my story is that Linux and Genomics grew, grew quickly and are affecting large changes. It's intersting to see that the rate of accumulation of sequence records in Genbank, the public sequence repository, is the same as the rate of increase of the number of web pages about linux indexed by Google.

COLLINS

I mentioned that science needs openness. Unfortunately there are forces in the daily world of science that go against this need. Who wants their ideas leaked out before they're published. If publishing ensures the survival of the scientist, hiding data and ideas make natural bedfellows.

In Genomics data sets are very large and very, very expensive to generate. It is unthinkable to imagine some corporate organization hoarding the sequence to the human genome. It is also unthinkable that the process of science should reward scientists for encouraging such practise.

As more and more open-access journals start up, and popular ones like the Public Library of Science, continue to grow, good science is predicated on sharing.

GENOMICS DATA

Genomics started when the web started. Or very nearly. Therefore, from the beginning, those working in the field have been exposed to the ease with which data can be shared. They benefited from the data and ideas of others and therefore published their own findings with FTP or a web server.

These findings are never extremely obvious though. Just because the human genome is sequenced, this doesn't mean that we know what it does. We have reasonable spotty knowledge but not a thorough understanding. Who knows, we may never figure out the "why" – but we can hope to know the "how".

DATA VIEWERS

The data is public and its therefore fitting that the data viewers should be platform agnostic. Nearly all of the public genomic data is accessible from some URL through your browser. You can point and click your way to any part of the human, mouse or rat genome sequence, for example, view the annotations and download sequence. You can provide your own annotations through a distributed annotation system.

This means that the cost of entry into genomics is simply the willingness to learn the background and material. And, of course, some old 386 that you can dust off and install Linux on.

COMMUNITY

The genomics community is widely networked. If you go to a genomics conference, you'll find that many people are coding Perl on wireless laptops during talks. I hope nobody is doing that here.

You may recognize some of these names from their contributions outside of genomics. Lincoln Stein, for example, wrote CGI.pm and GD.pm. I don't know anyone who hasn't at least tried one or both of these modules. Lincoln's article about how Perl saved the human genome project illustrates some of the points I'm trying to emphasize. Linux allows you to cheaply roll up your sleeves and get useful work done.

Jim Kent beat Celera to the human sequence assembly with his open source gigassembler.

Ewan plays a central role in the BioPerl bundle, a collection of Perl modules for manipulating and analyzing biological information.

SARS – LINUX IN ACTION

I'd like to use the SARS phenomenon as an example of how Linux was used in one case to do one thing very very quickly and very well. You may recall the SARS "epidemic" last year.

Our center was the first to assemble and publically release the SARS genome. All the analysis and assembly work was done on Linux workstations. The data was published and served from a Zope/Apache system.

OPEN DATA

As soon as we had it, the SARS genome was submitted to Genbank. By now, there are nearly 300 sequences from 150 different strains of SARS. There are over 1,600 SARS-related publications indexed by Pubmed.

If you wanted to see what the SARS genome looks like – you can. Download a 30kb text file from Genbank and you can have your very own copy of the virus.

SLASHDOT

At the time, SARS was receiving a lot of attention. Not much about it was known and people were scared. A large outbreak in Toronto diminished the city's tourism to the point where they had to drag Mick Jagger out of the closet and have him stage a concert to show that Toronto was safe. Well, at least from SARS.

You can see Duane here, the only employee from our Center who has a full color cover spread of a famous local chinese publication. He tells me that, although prolific, this kind of exposure didn't make him more popular with the ladies. Perhaps it's all the SARS plates he's holding.

PUBLIC

We got attention from the media and also from the public. Most people were very excited to participate in the discovery. A lot of labs took our data and continued the research process.

But... there's always someone out there that thinks you're crazy. This individual felt that we should not publish virus sequence on the internet. There has to be a better way. I don't know if there is – for now, I think we did it the right way.

LINUX AT GSC

I remember my first day on the job. Steve Jones, my boss, said: "We can't get the tape autoloader working well with the server." The server was a dual P2 beige box. The autoloader was an external DLT robot with 8 slots. The kernel was 2.0.36.

Admittedly, if we were to start a genome center today it would not be innovative to use Linux. I couldn't really stand here and say that we found this great operating system and it did everything and we used it and it worked. Ho-hum, right? However, if this were a Windows conference, and I said that we found this great OS from Redmond and used it to start a genome center people would think we're either innovative or crazy.

LINUX AT GSC 2

In 2000 Linux was different than today. The hardware was different. But both worked. I struggled with NFS problems, trying to export 400 GB to about 20 users hammering the disks. But it worked. We had a box what stayed up for

393 days – it did our NFS, NIS, mail and web. It continued to work well after it was the slowest machine on the network – just a dual 500 MHz Xeon.

LINUX AT GSC 3

We were starved for CPU power. You couldn't get a cheap fast CPU – unless you bought one or two of these overclocked beasts. They sat on a refrigeration unit and had their AMD CPUs cooled to -40C. Both polaris and kronos eventually broke down but they managed to assemble the human fingerprint map in record time. Kronos was one of the first 1GHz boxes on the block.

LINUX AT GSC 4

Our second generation machines were from VA Linux. Our first cluster platform was the IBM x330. These things were rock solid. Linux just ran and ran and ran on them. Things crashed but we knew why things crashed.

LINUX AT GSC 5

I can't overstate the importance of MySQL in our Center. It runs the entire LIMS system. We have wireless barcode scanners mounted on iPAQs that interface with the web-based database interface. Everything has a barcode on it. We print barcodes on network Zebra printers with the help of homegrown Perl-to-ZPL wrapper.

LINUX AT GSC 6

Our LIMS schema grew from about 20 tables to 200. Here's an example of how the equipment table interacts with other tables. There's five people coding Perl and Apache modules to deal with this database.

It stores everthing. Every plate in the lab, every solution, every step. We can look back and ask why the data was poorer on that day with that solution on that sequencer.

LINUX AT GSC 7

I was going to talk about some of the hardware we used, but there's just too much. Here's a sample of some of the fun rack mounted stuff we have in the server room.

Netapps are used for main NFS storage. We have about 50 TB right now on various disks, ranging from 70Gb SCSI disk trays hooked to Raidion controllers to 140 GB fibre channel disks hooked to Netapps. We have some blade servers, a lot of the IBM x330s, which were the first cluster work horses. More and more we're getting into opterons. There's a walk-in closet LTO2 robot with 400 slots.

We just unloaded a 96GB sunfire e2900 and two rack fulls of sun opterons. Huge memory machines like the sunfire are used to handle assemblies of large genomes, which routinely require 32GB of memory.

PLANS

This brings me to the end of the talk. I can stand here and talk about these things because a handful of people made some big plans. Plans to write a better operating system. Plans to develop robotics and sequencing technology. Plans to make these revolutionary ideas public.

There's a saying in science, "if you didn't write it down, you didn't do it." Well, here's another take on it: "if you don't share it, someone else will".

Whereeever go you, there will be people just as smart and smarter than you. If you think of it, they'll think of it too. Concepts will connect and make their way out into the world, there is no stopping revolutions.

The best thing to do, I think, is to share right away. Heck, before it's ready. In research you rarely get a chance to spit and polish anything. One finished project means another started. Newton maintained that he could see further because he stood on the shoulders of giants. The world needs more shoulders.

JOBS

Thanks.