

Canadian Bioinformatics Workshops

www.bioinformatics.ca



This page is available in the following languages:

Afrikaans বাংলা Azərbaycanca Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Español Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Malayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски (latinica) Sotho svenska
中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work



Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
[English](#) [French](#)

[Learn how to distribute your work using this licence](#)

Data Management

Asim Siddiqui
Bioinformatics Workshop
Next Generation Sequencing
26th July 2008

It's a new world

- Next gen sequencers generated Gb of data
- Architecture matters

Why should you care?

- Architecture matters
 - If you are choosing the compute resources for your lab, designing the correct system architecture is important to get the most use out of the systems
 - If you are using the compute resources of your lab, understanding the system architecture will help you get the most out of the system

What you won't learn.

- This lecture will provide you with a basic understanding of computer systems, but....
- Consult your local expert
- If you are the local expert, make friends at a large genome centre 😊

The basics

- CPU
- Disk space
- RAM
- Bandwidth

CPU

- The speed of your CPU determines how quickly it can process instructions
- Many bioinformatics operations fall into the “embarrassingly parallel” category
- Getting a results faster is as simple as adding more CPUs
- => clusters

How many CPUs?

- For current throughputs, you will need ~8 CPUs per sequencer to handle data rates
- 8 way boxes are relatively easy to come by.

RAM

- RAM is fast storage close to the CPU
- By loading data from disk to RAM, the CPU can execute instructions much more rapidly

How much RAM?

- Typical sizing is 2GB of RAM per CPU
- This works fine for most aligners
- Assemblers typically need much more RAM
- If you don't have enough RAM, the CPU will need to make use of the disk storage –
- When a computer has run out of RAM it is said to be “swapping”

Disk space

- Unlike RAM, information is retained after the machine is switched off
- Speed of access is slower than RAM
- Magnetic disks have a seek time and read time
 - Read data from a block is faster than seeking all over the disk to get the data
- Can be RAIDed to improve performance

How much disk space?

- An Illumina GA2 generates 5.35 GB per run (3 days)
- Including quality values and additional files results in 60GB per run
- Each machine will generate 7.3TB of data per year
 - Plus you will need to ~double that to store alignments and other data derived from the reads themselves
- Scaling

Space, or lack thereof, is the problem

- Disk space is probably the biggest problem today
- 60GB is for sequence data and quality values
- Should you store images?

To store or not to store?

- Storing images or their derivative, intensity values, is probably the biggest question at this time
 - SOLiD/Illumina generate >1TB per run
 - Helicos ~50TB per run
- Fridge vs. amortization value of machine time

Understanding bandwidth

- Bandwidth of a connection represents the maximum rate of transfer between two points
- Most commonly, we think of network bandwidth, but there is also bandwidth
 - between the disk and CPU
 - between a RAID array and the CPU
 - between the RAM and the CPU

More bandwidth

- Depending on the algorithm, the CPU will process data at a particular rate
- The trick is always max out the CPU's utilization
- Bandwidth to the CPU must be \geq the rate at which the CPU can process data

Even more bandwidth

- E.g. An aligner can process X reads per second on a single CPU at a data rate of Y bytes/sec
 - ~200 million reads in 10 hours
 - Each read 50 bases at 10 bytes per base
 - 2.7 MB/sec
- Design 100TB storage and connect it to a CPU resource
- Design bandwidth to be 10 Mb/sec – plenty of spare bandwidth
- Now we want to complete the job in an hour and get permission to buy 10 more CPUs – great!

Not so great

- For the 10 CPUs to run at maximum speed, they need to be supplied data at 27MB/sec
- Our bandwidth is 10MB/sec
- Therefore, no matter how many CPUs we buy, the job will never run faster than ~ 2.5 hours

Balancing it all

- The best balance of compute resources is application dependant
- Aligners may require a different balance than assembler (or other algorithms)
- Decisions
 - If limited resource, design system to deal with the biggest bottleneck
 - Ideally, have different systems for different parts of the pipeline

Backing it up

- Where to start....
 - Who's backing up their data today?
- Back up to active disk is probably the easiest
- Backup to tape expensive
 - Person time
 - Need to refresh tapes
 - Slow
- Active disk can't be taken offsite
- No great solutions out there... Except perhaps SEP machine...

Clouds

- Clouds, such as Amazon EC2, are emerging as a viable alternative to owning your own resources
 - Remote disk storage
 - Remote CPU
 - Pay as you go
- This could be a good option for a small lab

LIMS

- Generating all that data is no good if you don't know where you've put it
- LIMS provide a mechanism for keeping track of all of the data
- Metadata (i.e. data about data) related to the experiments is stored in a database
- The database stores the location of the data files
- Tracking may be paper based or bar code based
- Essential for a centre running lots of expts

LIMS and other s/w

- Build in-house and off-the-shelf
 - Commercial or free solutions
- Is there a solution that meets your needs?
 - If missing some requirements, will the company/lab modify their s/w for your needs?
- How do you want to spend your time?

Emerging Standards

- Sequence
- Alignment
- Best practises
- Why are standards important?
 - Hint: wouldn't you just like to focus on the science?
- “If you build it, will they come?”