Informatics Challenges
Next–generation sequencing

HEMANT KELKAR

Center for Bioinformatics

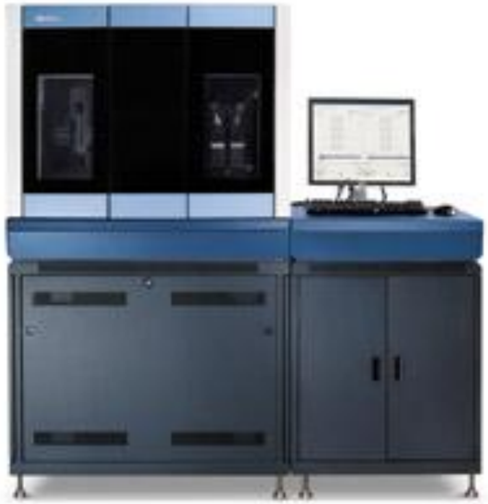
UNC-Chapel Hill, NC 27599

hkelkar@unc.edu

http://bioinformatics.unc.edu

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Now-Generation Sequencing



2010

- > 1,000,000,000 bases (1 GB) per hour

Challenges ….

1. Data volume

Challenge for the end-user…

- 3.6G Apr 27 20:13 s_2_sequence.txt
- 6.0G Apr 27 20:20 s_3_sequence.txt
- 6.1G Apr 27 20:21 s_1_sequence.txt
- 5.0G Apr 27 20:21 s_5_sequence.txt
- 6.0G Apr 27 20:25 s_4_sequence.txt
- 4.9G Apr 27 20:30 s_6_sequence.txt
- 5.5G Apr 27 20:42 s_8_sequence.txt
- 5.6G Apr 27 20:42 s_7_sequence.txt

Would you like fries with that …

- 5.0G Apr 27 20:21 s_5_sequence.txt
- 4.2G Apr 27 20:40 s_5_eland_extended.txt
- 3.6G Apr 27 22:19 s_5_sorted.txt

- 13 GB of data **for one sample**

HiSeq 2000

Changes loom in the data landscape .. yet again*

- Image Data – 32 TB (not kept)
- Intensity Data – 2 TB (may want to keep)
- Base Call/Quality data – 250 GB
- Alignment Output – 6 TB (1.2 TB if intermediates removed)

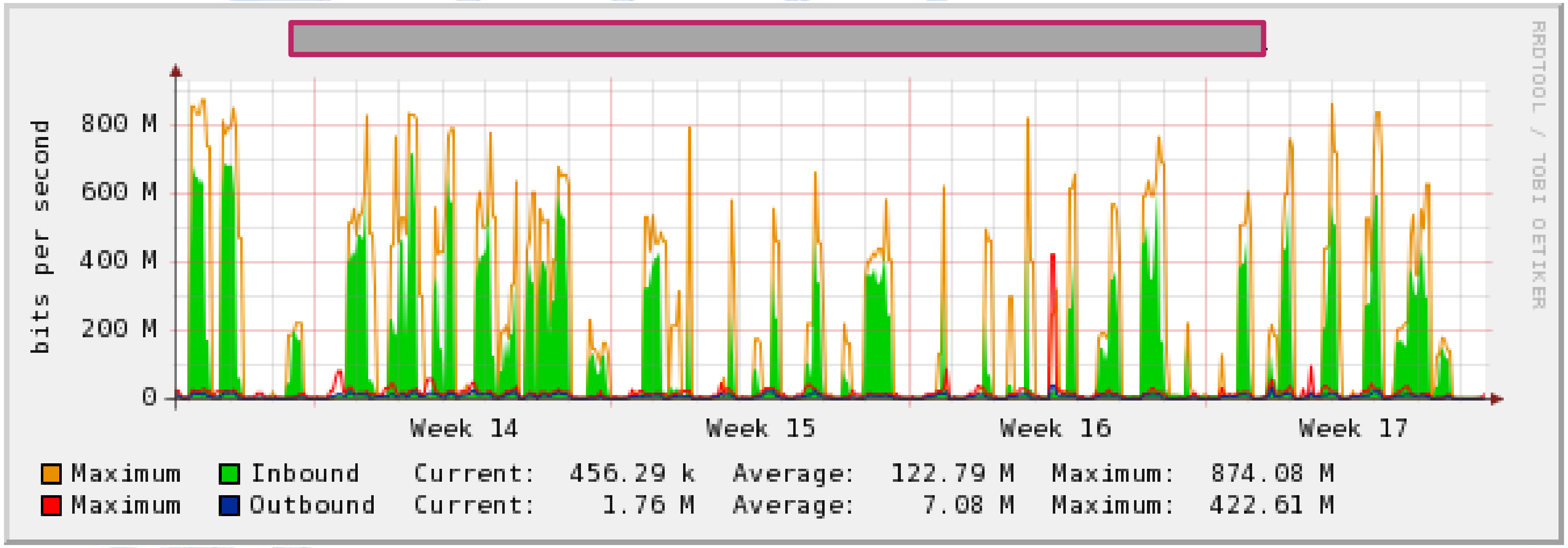
* Numbers from Illumina

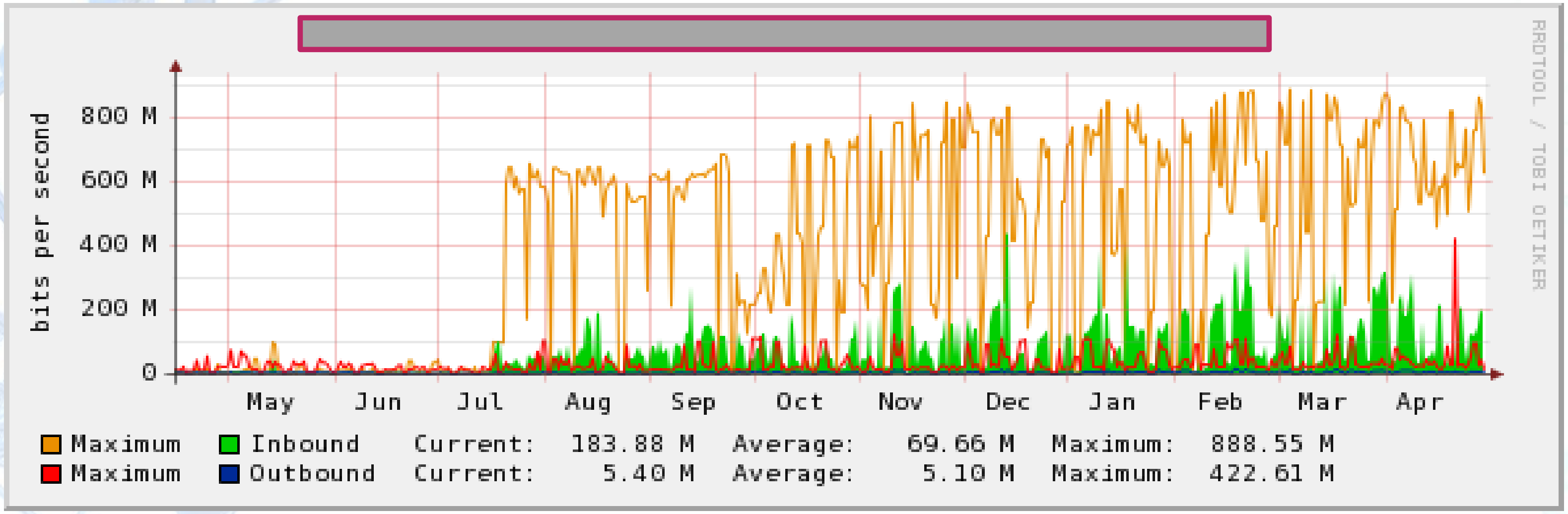
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Challenges ….

1. Data volume
2. Network bandwidth

Network under siege


[redacted]


[redacted]

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Challenges ….

1. Data volume
2. Network bandwidth → cross-mounting partitions across WAN/VLANs may not be a viable option
3. Time

Time

- Time to download/copy/delete/process
- Copying data takes significant amount of time

(1-2 GB/min)

- Time to align/assemble 20-30 million reads (e.g. few hours to human genome)
- Hard to do truly parallel software
- I/O bottlenecks

Challenges ….

1. Data volume
2. Network bandwidth
3. Time
4. Hardware

Hardware

Client side :

- Storage may not be a big issue
- Do remember to have a plan for data backup
- 64-bit OS (Windows, Mac OS X, Linux)
- A dedicated workstation .. if possible

Facility end :

- Storage is a big issue – how much/how long/long term
- Data backup strategy (tape, archival disk storage, ILM)
- LIMS
- Data release mechanism

Challenges ….

1. Data volume
2. Network bandwidth
3. Time
4. Hardware
5. Sharing/Publishing

Sharing data

- NCBI GEO accepts HT sequence data
- NLM – SRA (sequence read archive)
- SRF file format developed at Sanger http://sequenceread.sf.net
- Web .. Bandwidth
- Cloud computing .. Specific storage formats

Challenges ….

1. Data volume
2. Network bandwidth
3. Time
4. Hardware
5. Sharing/Publishing
6. Personnel