Storage, Bandwith and Sequence Alignment

Lou Shaoke

Department of Molecular Biophysics and Biochemistry

loushaoke@gmail.com

June 2, 2015

Yale

Task list

Outline

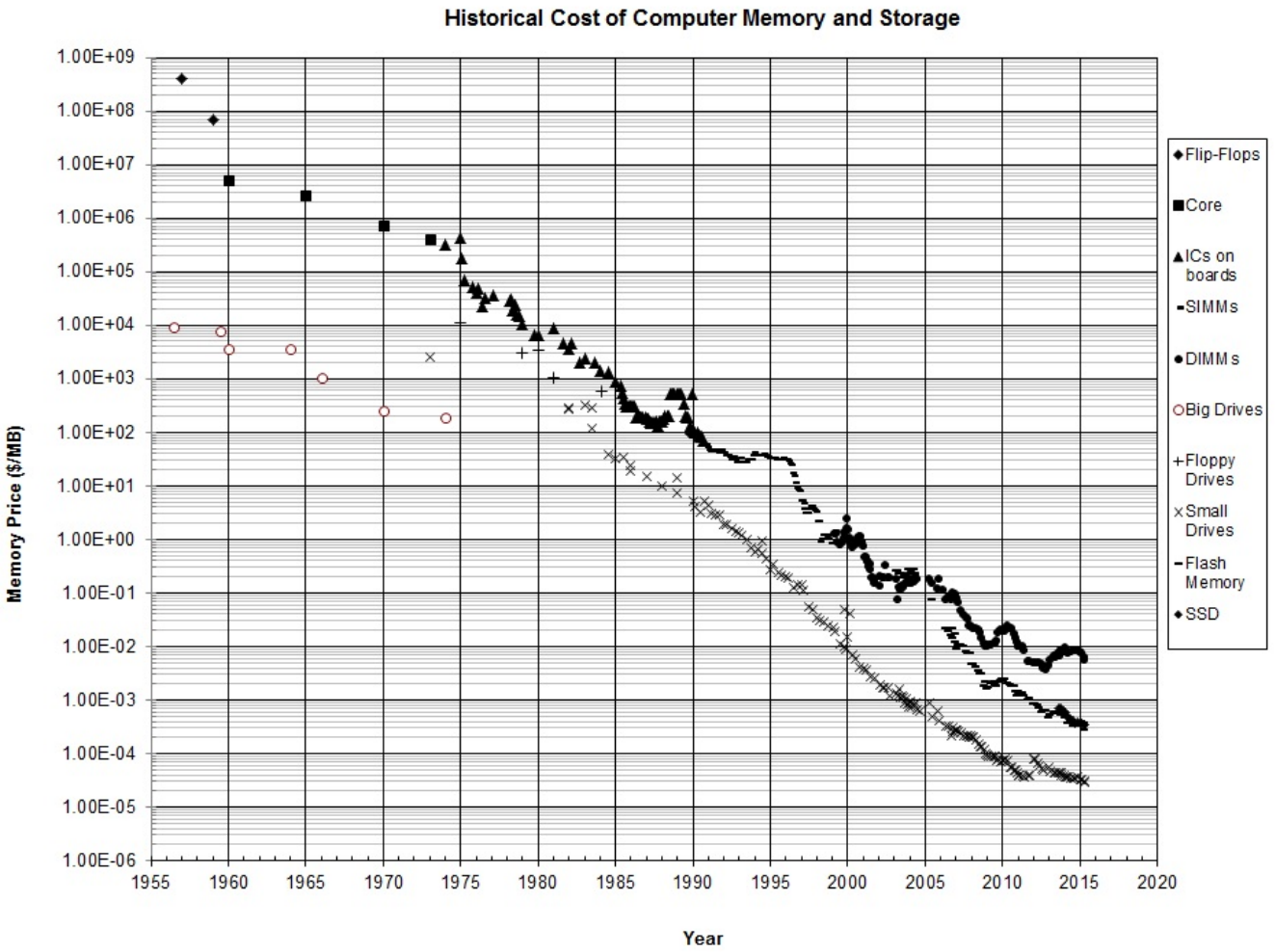
The computational component of sequencing and importance of scalable storage and search technologies. How does the analysis component scale? Alignment algorithms (BLAT, BWA, etc)? How is the differential scaling related to the rise of cloud computing and other changes to computing paradigms?

JZ and SKL think about:
-Storage and bandwidth graph -Peta store vs. giga store -PCAWG dataset size and theoretical download time $/year invested into pharma pipelines
Alignment algorithm improvement[1]
-NW
-FASTA
-BLAST
-BLAT
-BWA
comparison to tree building

[1] Heng Li and Nils Homer. "A survey of sequence alignment algorithms for next-generation sequencing." In: *Briefings in bioinformatics* 11.5 (Sept. 2010), pp. 473–83. ISSN: 1477-4054. DOI: 10.1093/bib/bbq015. URL: http://bib.oxfordjournals.org/content/11/5/473.full.

Disk cost

Hard disk cost(`http://www.jcmit.com/diskprice.htm`)



Price reduce 10^{13} in 60 Years, around 0.6 (or half) each year. It makes possible for cloud storage.

Yale

Bandwidth

Broadband community annual report 2014:

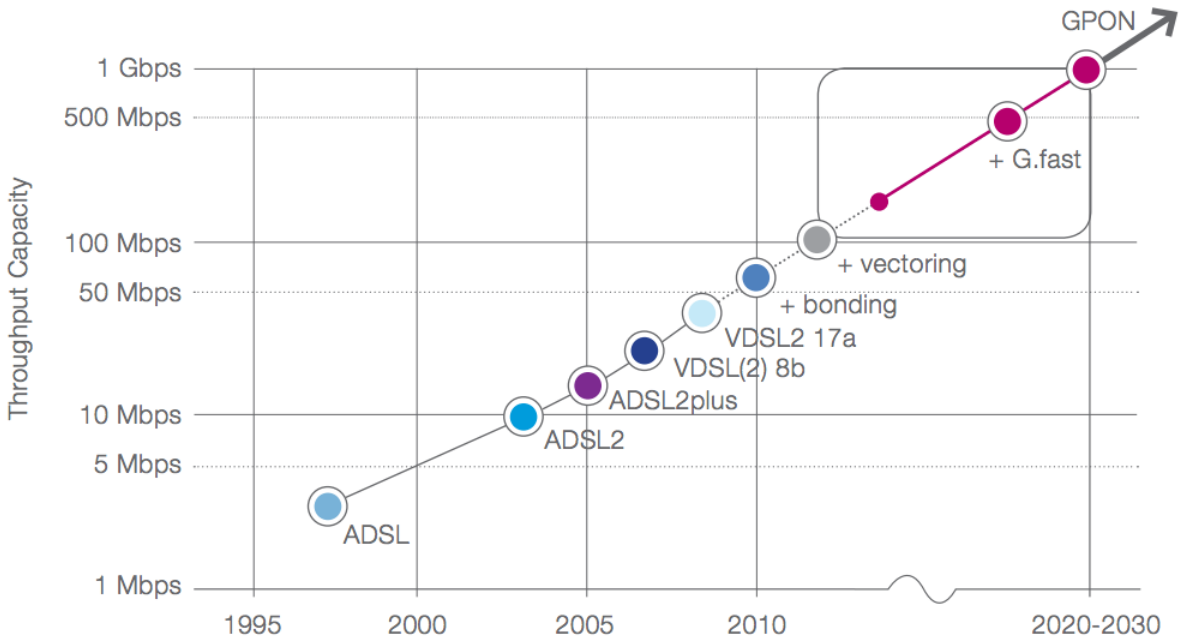


Figure 3: Growth in Speeds for Fixed and Mobile Technologies

The evolution of copper – bridging the gap between xDSL and fibre speeds (top); The shift in subscriptions towards mobile technologies with higher speeds (bottom).

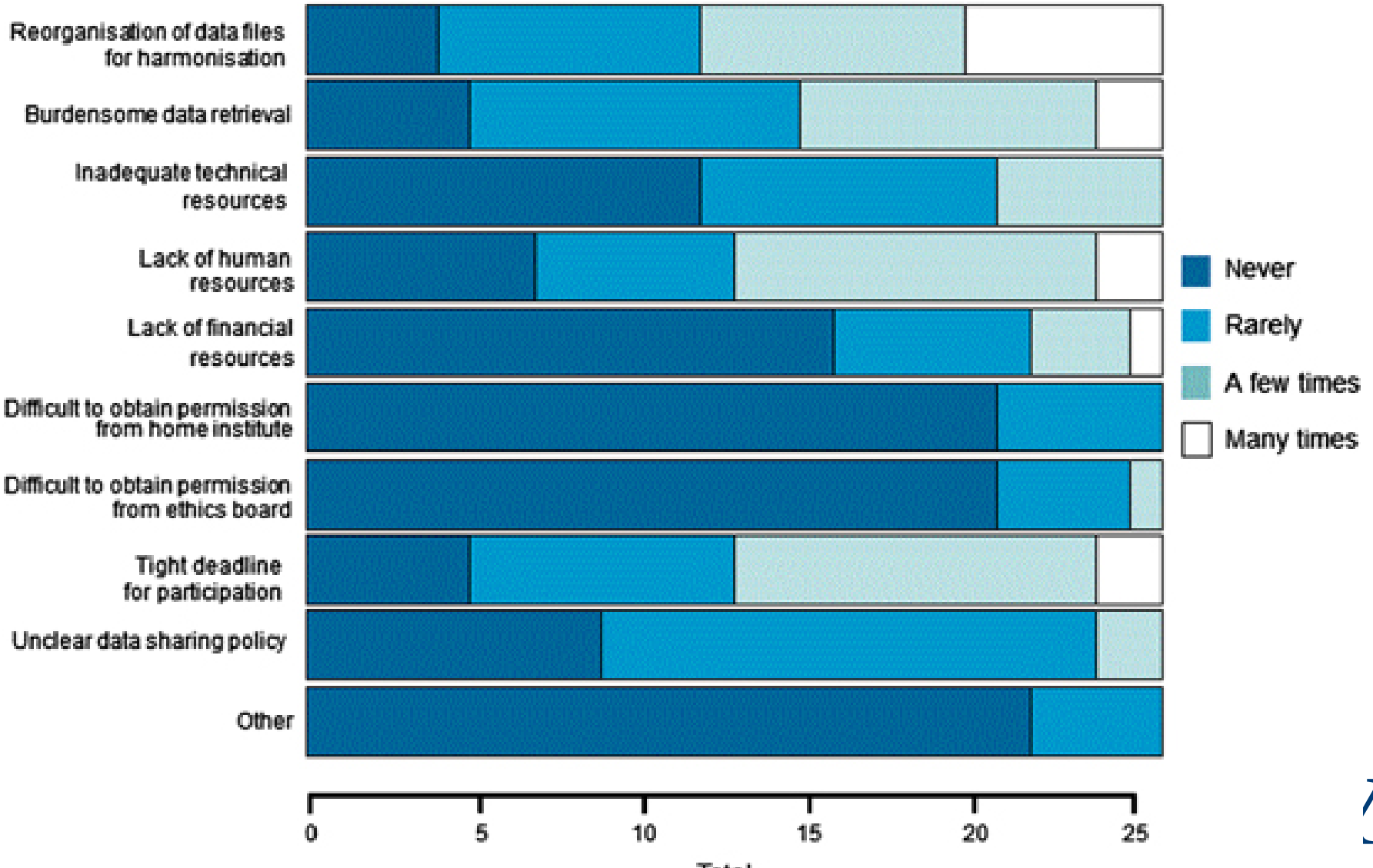
Source: Alcatel Lucent (top chart), Ericsson Mobility Report, June 2014 (bottom chart).

Now the bandwidth is around 200Mbps, varies in different nations (FCC 4th report).

Yale

Bandwidth

Bandwidth is one of the big hurdles (data organization and retrieval, human resources, Budin-Ljøsne et al. 2014):



Estimation to download/upload data to AWS Virginia

Site	MB/s to Virginia	Genome Alignments per Day	Variant Calling per Day	Days to Transfer 100TB
cghub.ucsc.edu (Santa Cruz)	18.26	2.57	5.14	63.39
gtrepo-dkfz.annailabs.com (Heidelberg)	28.83	4.05	8.11	40.14
gtrepo-ebi.annailabs.com (London)	23.00	3.23	6.47	50.31
gtrepo-etri.annailabs.com (Seoul)	0.00	0.00	0.00	4740740.74
gtrepo-osdc-icgc.annailabs.com (Chicago)	35.78	5.03	10.06	32.34
gtrepo-riken.annailabs.com (Tokyo)	49.08	6.90	13.80	23.58

* Per day estimates are based on a single upload/download client transferring 300GB/genome each for unaligned (download) and aligned (uploaded) files (600GB total).

Yale

Sequence alignment

NW, SW, FASTA and BLAST: once upon a BLAST (Filion n.d.) and A survey of sequence alignment algorithms for next-generation sequencing (Li and Homer 2010).

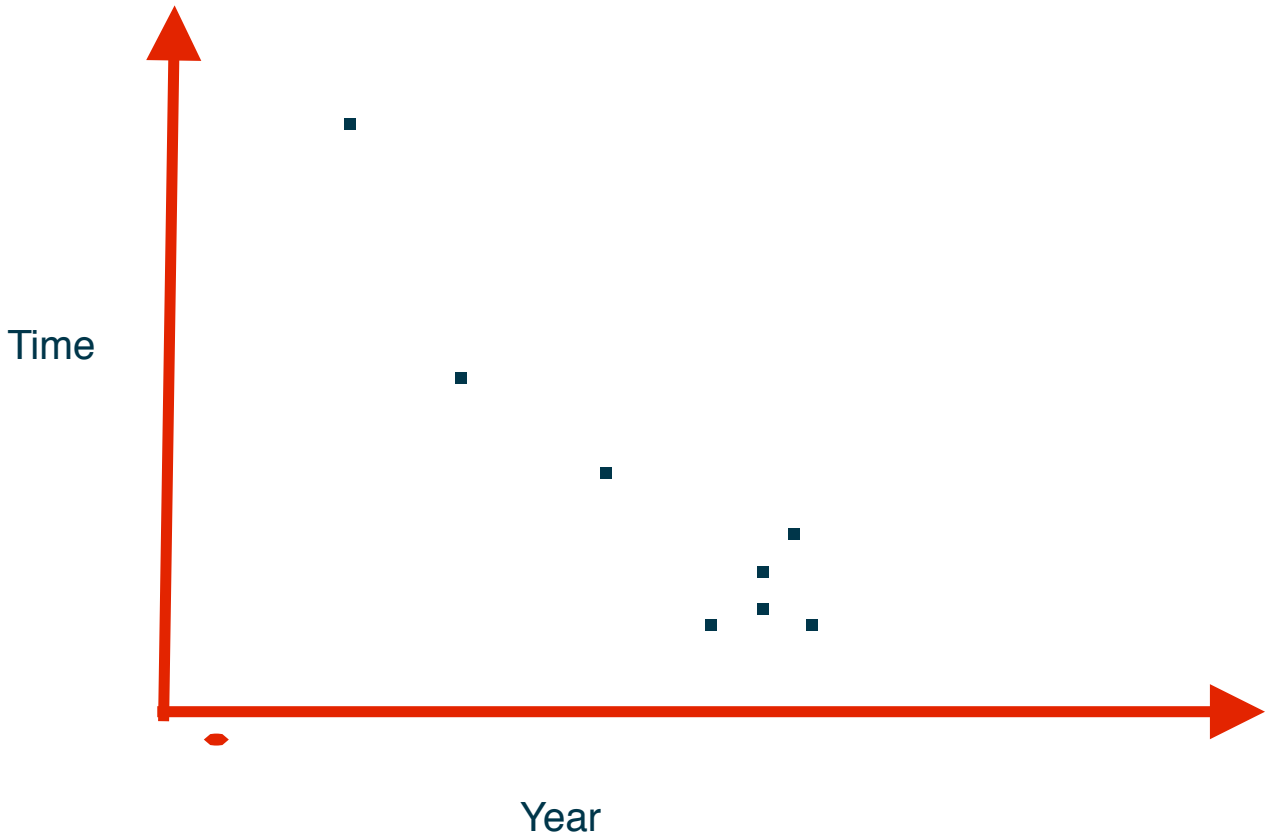
Name	Year	Algorithm	attrib	Reference
Needleman-Wunsch	1970	DP	pairwise alignment	Needleman and Wuns
Smith-Waterman	1981	DP	pairwise alignment	Smith and Waterma
FASTA	1985	Heuristic word	database search	Lipman and Pearso
BLAST	1990	K-mer	database search	Altschul et al. 1
BLAT	2002	K-mer	database search, NGS	Kent 2002
BWA	2009	BWT	NGS	Li and Durbin 2

The most important feature for tools in NGS era is the use of auxiliary structure: hash-table, suffix array, FM-index, BWT and compressed suffix array et cetera.

Data management portal/pipeline: biomart, ucsc genome brower, Galaxy and seqware et cetera.

Yale

Figure like this?



Yale

References I

S F Altschul et al. "Basic local alignment search tool." In: *Journal of molecular biology* 215.3 (Oct. 1990), pp. 403–10. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2. URL: http://www.ncbi.nlm.nih.gov/pubmed/2231712.

Isabelle Budin-Ljøsne et al. "Data sharing in large research consortia: experiences and recommendations from ENGAGE." In: *European journal of human genetics : EJHG* 22.3 (Mar. 2014), pp. 317–21. ISSN: 1476-5438. DOI: 10.1038/ejhg.2013.131. URL: http://dx.doi.org/10.1038/ejhg.2013.131.

Guillaume Filion. *Once upon a BLAST*. URL: http://blog.thegrandlocus.com/2014/06/once-upon-a-blast.

W. J. Kent. "BLAT—The BLAST-Like Alignment Tool". In: *Genome Research* 12.4 (Mar. 2002), pp. 656–664. ISSN: 1088-9051. DOI: 10.1101/gr.229202. URL: http://genome.cshlp.org/content/12/4/656.full.

Yale

References II

Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform." In: *Bioinformatics (Oxford, England)* 25.14 (July 2009), pp. 1754–60. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp324. URL: http://bioinformatics.oxfordjournals.org/content/25/14/1754.abstract?ijkey=dece3ec9182f6a8ddc619c4fb33135496d1f8d9b%5C&keytype2=tf%5C_ipsecsha.

Heng Li and Nils Homer. "A survey of sequence alignment algorithms for next-generation sequencing." In: *Briefings in bioinformatics* 11.5 (Sept. 2010), pp. 473–83. ISSN: 1477-4054. DOI: 10.1093/bib/bbq015. URL: http://bib.oxfordjournals.org/content/11/5/473.full.

D J Lipman and W R Pearson. "Rapid and sensitive protein similarity searches." In: *Science (New York, N.Y.)* 227.4693 (Mar. 1985), pp. 1435–41. ISSN: 0036-8075. URL: http://www.ncbi.nlm.nih.gov/pubmed/2983426.

S B Needleman and C D Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." In: *Journal of molecular biology* 48.3 (Mar. 1970), pp. 443–53. ISSN: 0022-2836. URL: http://www.ncbi.nlm.nih.gov/pubmed/5420325.

Yale

References III

T F Smith and M S Waterman. "Identification of common molecular subsequences." In: *Journal of molecular biology* 147.1 (Mar. 1981), pp. 195–7. ISSN: 0022-2836. URL: http://www.ncbi.nlm.nih.gov/pubmed/7265238.

Yale