Hadoop-BAM and SeqPig

Keijo Heljanko[1], André Schumacher[1,2], Ridvan Döngelci[1], Luca Pireddu[3], Matti Niemenmaa[1], Aleksi Kallio[4], Eija Korpelainen[4], and Gianluigi Zanetti[3]

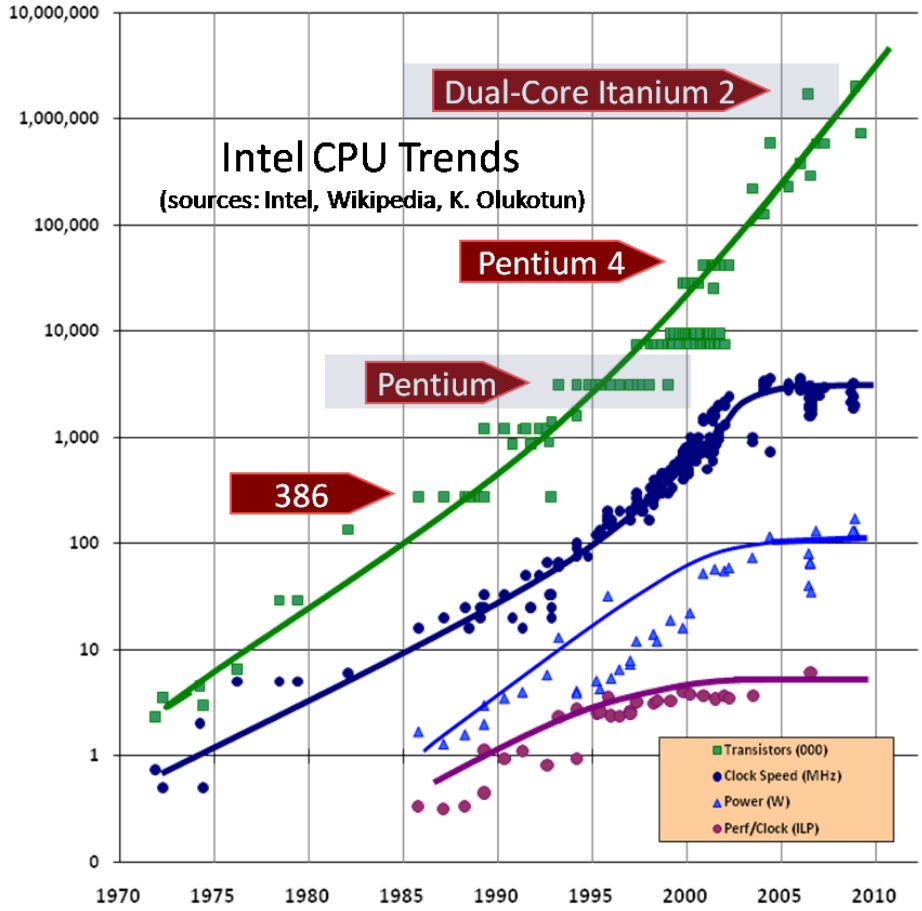
[1] Department of Computer Science and Engineering and
Helsinki Institute for Information Technology HIIT
School of Science, Aalto University
firstname.lastname@aalto.fi
[2] International Computer Science Institute, Berkeley, CA, USA
[3] CRS4 — Center for Advanced Studies, Research and Development in Sardinia, Italy
[4] CSC — IT Center for Science

20.1-2015

Next Generation Sequencing and Big Data

- The amount of NGS data worldwide is predicted to double every 5 months
  - This growth is much faster than Moore's law for the growth rate of computing (historically transistor counts have doubled every 18-24 months), Kryder's law for the growth of storage capacity (historically doubling approx every 13 months), and Butter's law for growth in optical communications bandwidth (historically doubling approx every 9 months)
- Without increased expenditure in distributed computing methods genomics research will hit computational limits

No Processor Clock Speed Increases Ahead

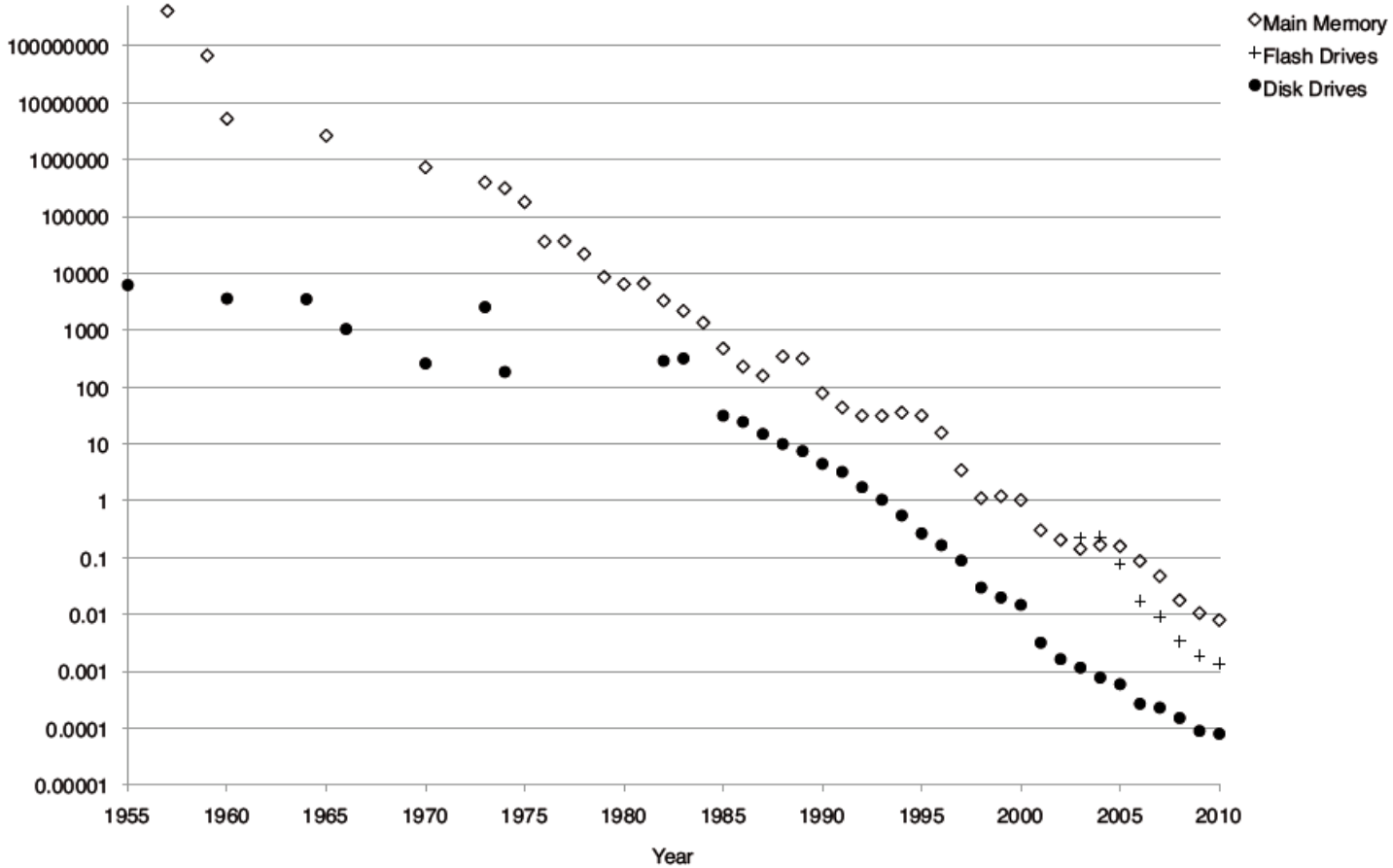


- Herb Sutter: The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software. Dr. Dobb's Journal, 30(3), March 2005 (updated graph in August 2009).

Implications of the End of Free Lunch

- The clock speeds of microprocessors are not going to improve much in the foreseeable future
  - The efficiency gains in single threaded performance are going to be only moderate
- The number of transistors in a microprocessor is still growing at a high rate
  - One of the main uses of transistors has been to increase the number of computing cores the processor has
  - The number of cores in a low end workstation (as those employed in large scale datacenters) is going to keep on steadily growing
- Programming models need to change to efficiently exploit all the available concurrency - scalability to high number of cores/processors will need to be a major focus

Tape is Dead, Disk is Tape, RAM locality is King



- Trends of RAM, SSD, and HDD prices. From: H. Plattner and A. Zeier: In-Memory Data Management: An Inflection Point for Enterprise Applications

Tape is Dead, Disk is Tape, RAM locality is King

- RAM (and SSDs) are radically faster than HDDs: One should use RAM/SSDs whenever possible
- RAM is roughly the same price as HDDs were a decade earlier
  - Workloads that were viable with hard disks a decade ago are now viable in RAM
  - One should only use hard disk based storage for datasets that are not yet economically viable in RAM (or SSD)
  - In memory distributed filesystems such as Tachyon are needed for temp files!
  - The Big Data applications (HDD based massive storage) should consist of applications that were not economically feasible a decade ago using HDDs

Hadoop - Linux of Big Data

- Hadoop = Open Source Distributed Operating System Distribution for Big Data
  - Based on Google architecture design
  - Cheap commodity hardware for storage
  - Fault tolerant distributed filesystems: HDFS, Tachyon
  - Batch processing systems: Hadoop MapReduce, Apache Hive, and Apache Pig (HDD); Apache Spark (RAM)
  - Parallel SQL implementations for analytics: Apache Hive, Cloudera Impala, Apache Shark, Facebook Presto
  - Fault tolerant distributed database: HBase
  - Distributed machine learning libraries, text indexing & search, etc.
  - Project Web page: `http://hadoop.apache.org/`
- Hadoop MapReduce is just one example application on top of the Hadoop Open Source distribution!

Commercial Hadoop Support

- Cloudera: Probably the largest Hadoop distributor, partially owned by Intel (740 million USD investment for 18% share). Available from:
  `http://www.cloudera.com/`
- Hortonworks: Yahoo! spin-off from their large Hadoop development team:
  `http://www.hortonworks.com/`
- MapR: A rewrite of much of Apache Hadoop in C++, including a new filesystem. API-compatible with Apache Hadoop.
  `http://www.mapr.com/`

Hadoop-BAM

- A library to interface NGS data formats with both Hadoop and Spark
- Includes tools for e.g., sorting of reads, as needed by merging results of parallel read aligners
- Supported fileformats: BAM, SAM, FASTQ, FASTA, QSEQ, BCF, and VCF
- Some fileformats like BAM notoriously badly designed for parallel processing
- Version 7.0 of the `hadoop-BAM` released:
  `http://sourceforge.net/projects/hadoop-bam/`
- 2700+ Downloads of the library
- Niemenmaa, M., Kallio, A., Schumacher, A., Klemelä, P., Korpelainen, E., and Heljanko, K.: Hadoop-BAM: Directly Manipulating Next Generation Sequencing Data in the Cloud. Bioinformatics 28(6):876-877, 2012. (`http://dx.doi.org/10.1093/bioinformatics/bts054`).

SeqPig

- Parallel scripting language for NGS data sets based on the Apache Pig language
- Compiles into Java, executed by Hadoop MapReduce
- SQL-like functionality with helper functions for NGS data: Filtering data, computing aggregate statistics, doing joins
- Supported fileformats: BAM, SAM, FASTQ, QSEQ, and FASTA
- Schumacher, A., Pireddu, L., Niemenmaa, M., Kallio, A., Korpelainen, E., Zanetti, G., and Heljanko, K.: SeqPig: Simple and scalable scripting for large sequencing data sets in Hadoop. Bioinformatics 30 (1): 119-120, 2014. (dx.doi.org/10.1093/bioinformatics/btt601.)
- See also supplement:
  http://bioinformatics.oxfordjournals.org/content/suppl/2013/10/17/btt601.DC1/supplement.pdf

SeqPig Use Case Examples

- Automatically parallelizing Pig example scripts for:
  - File format conversion
  - Filtering out unmapped reads and PCR or optical duplicates
  - Filtering out reads with low mapping quality
  - Filtering by regions (samtools syntax)
  - Sorting BAM files
  - Computing read coverage
  - Computing base frequencies (counts) for each reference coordinate
  - Pileup
  - Collecting read-mapping-quality statistics
  - Collecting per-base statistics of reads
  - . . .

Scalability of SeqPig

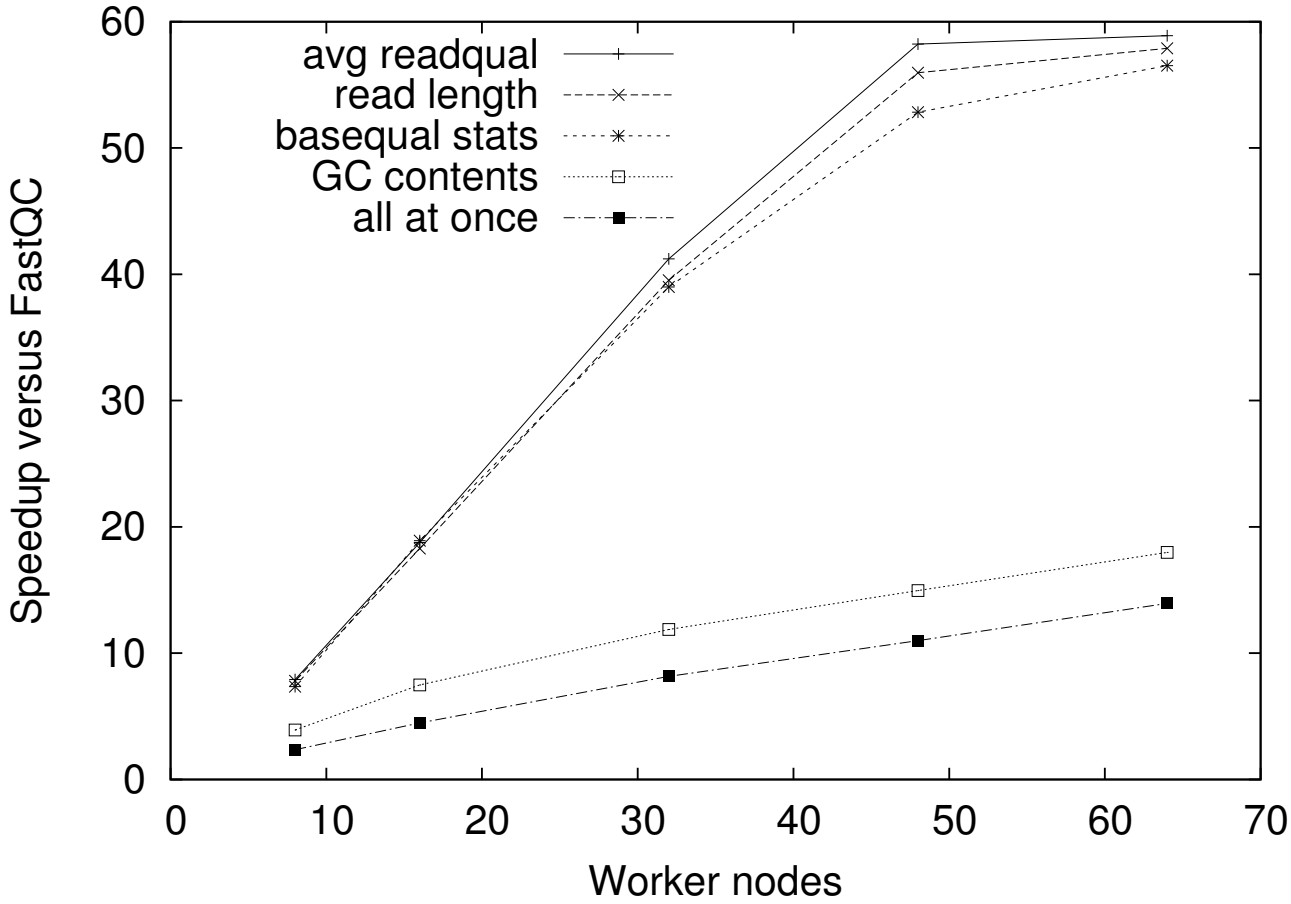


Figure: Scalabilty of SeqPig vs sequential FastQC. Computing statistics on 61.4 GB input file with up to 63 computer Hadoop cluster

SeqPig Benefits and Drawbacks

- Benefits:
  - Automatic parallelization of data processing scripts
  - Easy to learn scripting language with full power of MapReduce
  - Most scripts are at most tens of lines of code vs. hundreds to thousands of lines of Java
  - Also allows calling back user defined functions written in Java/Python
  - Implements SQL like functionality
- Drawbacks:
  - MapReduce has 10+ second startup delay: No for interactive use
  - A specialized language instead of a standard like SQL

Move from BAM files to SQL Data Warehouse

- A proper data warehouse system can
  - Very efficiently evaluate parallel queries over Petabytes of data
  - Allows for efficient compression and indexing
  - Allows to ride on the Hadoop+Spark software investments
- Many new analytics SQL implementations on top of Hadoop designed for handling Petabyte-class datasets
  - Apache Hive
  - Cloudera Impala
  - Spark SQL
  - Presto from Facebook

  We have interfaced Hive with NGS file formats using Hadoop-BAM, see Master's Thesis: Matti Niemenmaa: Analysing sequencing data in Hadoop: The road to interactivity via SQL, 2013.

Future plans

- Porting all our tools over to Spark from Hadoop
- Parallel variant detection using “black box” variant callers needs to be integrated with the pipeline - Work in progress
- Releasing a tool with SeqPig like functionality on a parallel SQL Data Warehouse
- Interactive ad-hoc queries of Big Data in Genomics
- In memory-filesystem and object storage integration
- We are open to working with you to help with your Big Data problems!