Benchmarking of Data Integration Workloads

Anthony Cabrera, Clayton Faber, Kyle Cepeda, Robert Derber, Cooper Epstein, Jason Zhang, Ron Cytron, and Roger Chamberlain

Dept. of Computer Science and Engineering, Washington University in St. Louis

CNS-1527510

Introduction

- Big data comes in many formats
- It often needs conversion, which can be time-consuming
- We are building a benchmark suite of data transformation applications

Data Transformations

Compositions of the following:

- Parsing
  - CSV, FITS, FASTA, FASTQ, SAM, UNIPEN, IDX3-UBYTE, Optdigits, etc.
- Cleansing
  - Detecting outliers, limit testing, etc.
- Transforms
  - Normalization, fixed-point → float, EBCDIC → ASCII, vector → pixel, formatting, (non-linear) scaling, etc.
- Aggregations
  - Summation, Histogram, etc.

Source Disciplines

Selected from:

- Computational Biology
  - Genomic and proteomic sequence data
- Astrophysics
  - Image processing
- IoT
  - Measuring the physical world
  - Large graphs
- Enterprise
  - Migration from mainframe to the cloud

Characterization

- Execution time
  - Expect *O(n)* – *n* is size of input
- Working set size
  - Expect *O(1)* – small relative to *n*
- Locality
  - Expect high spatial and low temporal locality
- Instruction mix (static and dynamic)
  - Expect to vary across applications
- Branch predictability
  - Expect to vary across applications
- Parallelizability
  - Often embarrassingly parallel across records

Some of the applications in the benchmark suite:

CSV Transformations

Taxi Service Trajectory

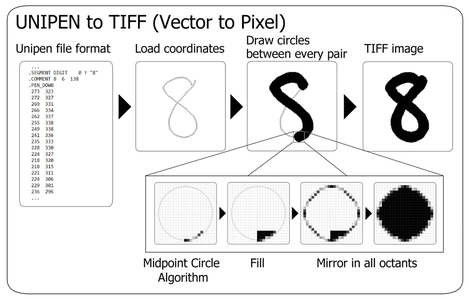
Generalized Format: [id],[lattitude],[longitude]

Go!Track GPS Trajectories

GeoLife GPS Trajectories

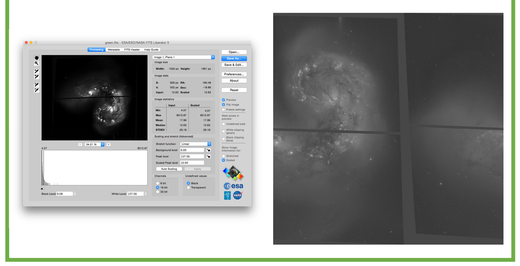
ECML PKDD 2015 Data Set

UNIPEN to TIFF (Vector to Pixel)



Astronomical Data Transformations

FITS format → TIFF



Computational Biology

Supported Input Formats:

- FASTA
- Multi-FASTA
- FASTQ
- SAM

Supported Output Formats:

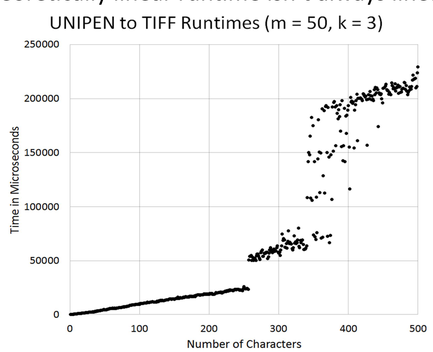
- 2 bits/base
- 4 bits/base

2 Bit Binary Transformation

Character	Binary Equivalent	2-Bit Sequence
A	0100 0001	00
C	0100 0011	01
T	0101 0100	10
G	0100 0111	11

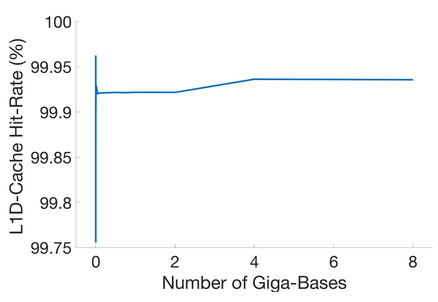
Some initial characterization measurements:

Theoretically linear runtime isn't always linear!



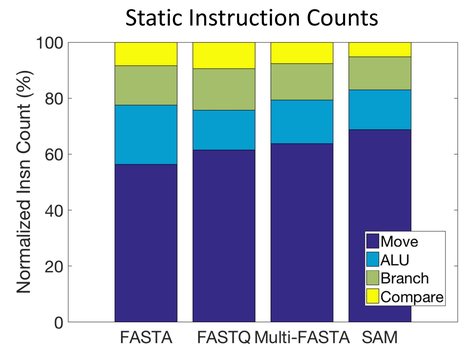
Genomics (FASTA → 2-bits/base)

Measuring Locality

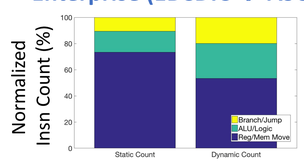


Genomics

Static Instruction Counts



Enterprise (EBCDIC → ASCII)



Raw Data Sources

https://archive.ics.uci.edu/ml/datasets/Taxi+Service+Trajectory+-+Prediction+Challenge,+ECML+PKDD+2015
https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits
http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits
http://www.geolink.pt/ecmlpkdd2015-challenge/
https://www.microsoft.com/en-us/download/details.aspx?id=52367
https://archive.ics.uci.edu/ml/datasets/GPS+Trajectories
https://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf
https://samtools.github.io/hts-specs/SAMv1.pdf
http://vann.lecun.com/exdb/mnist/
http://www.mersenneforum.org/showthread.php?t=1955|