DISTRIBUTED TILE PROCESSING

WITH

GEOTRELLIS AND SPARK

Rob Emanuele / @lossyrob

THE CHALLENGE

HOW DO WE WORK WITH **VERY** LARGE RASTER DATA?

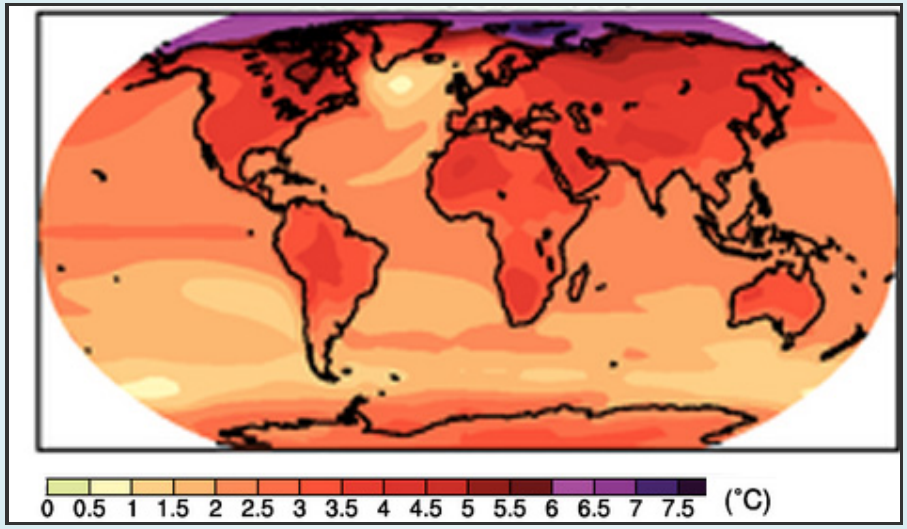
SPECIFICALLY...

HOW DO WE WORK WITH THE
NASA NEX DOWN-SAMPLED CLIMATE PROJECTIONS (NEX-DCP30)
OPEN DATA SET?

WHAT IS NEX CLIMATE PROJECTION DATA?

GLOBAL CIRCULATION MODELS

Models for predicting world temperature and precipitation.



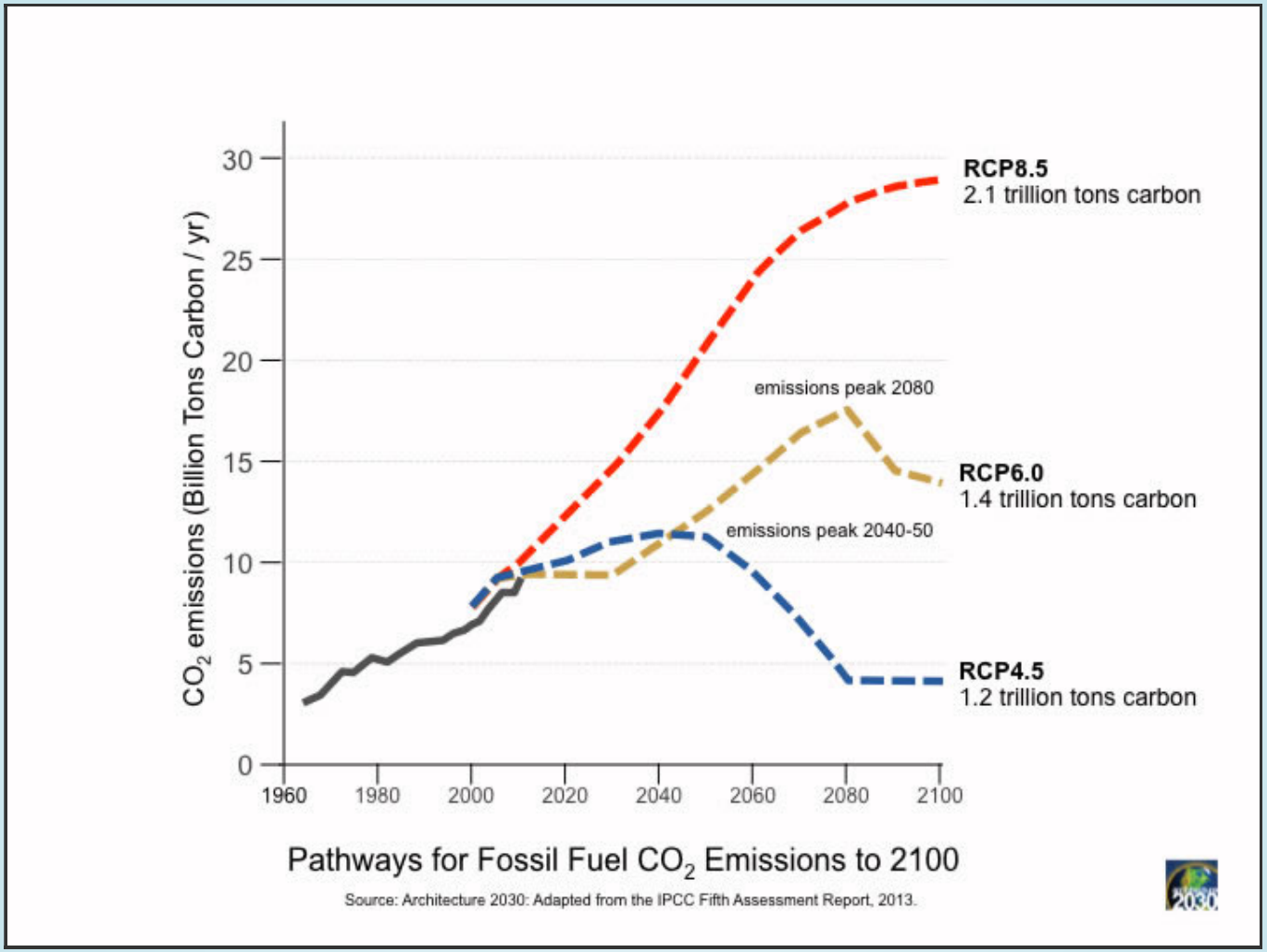
IPCC ASSESSMENT REPORT

- IPCC = Intergovernmental Panel on Climate Change
- Assessment Report 5 (AR5) published in 2014.
- More than 800 authors

3 KEY CATEGORIES:

- **MODEL**
  - **33** different models
  - Model Ensembling
- **DATASET**
  - Temperature MAX
  - Temperature MIN
  - Precipitation
- **SCENARIO**
  - Historical
  - Future RCPs

REPRESENTATIVE CONCENTRATION PATHWAYS



NEX DOWN-SAMPLED DATA

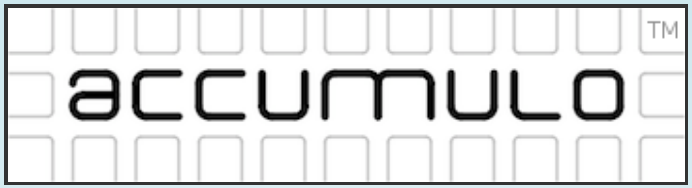
- Monthly data over conterminous US
  - Historical from 1950 - 2006
  - 4 RCP scenarios from 2006 - 2099
- 8190 netCDF files on S3 - s3://nasanex/NEX-DCP30
- **15.3** TB in compressed GeoTiff tiles.
- RCP 8.5, max for datatype/model combo: **90.92 GB**

OUR WORKFLOW FOR PROCESSING NEX DATA

THE TOOLS

- Scala library for doing all things geospatial.
- framework for doing distributed raster processing on Akka and Spark.
- Includes local, zonal, focal, and global operations on rasters.
- Currently in incubation at

- Fast and general engine for large-scale data processing
- Does things Hadoop doesn't, like cache intermediate results in memory.
- Written in Scala!
- Also has bindings for Python and Java

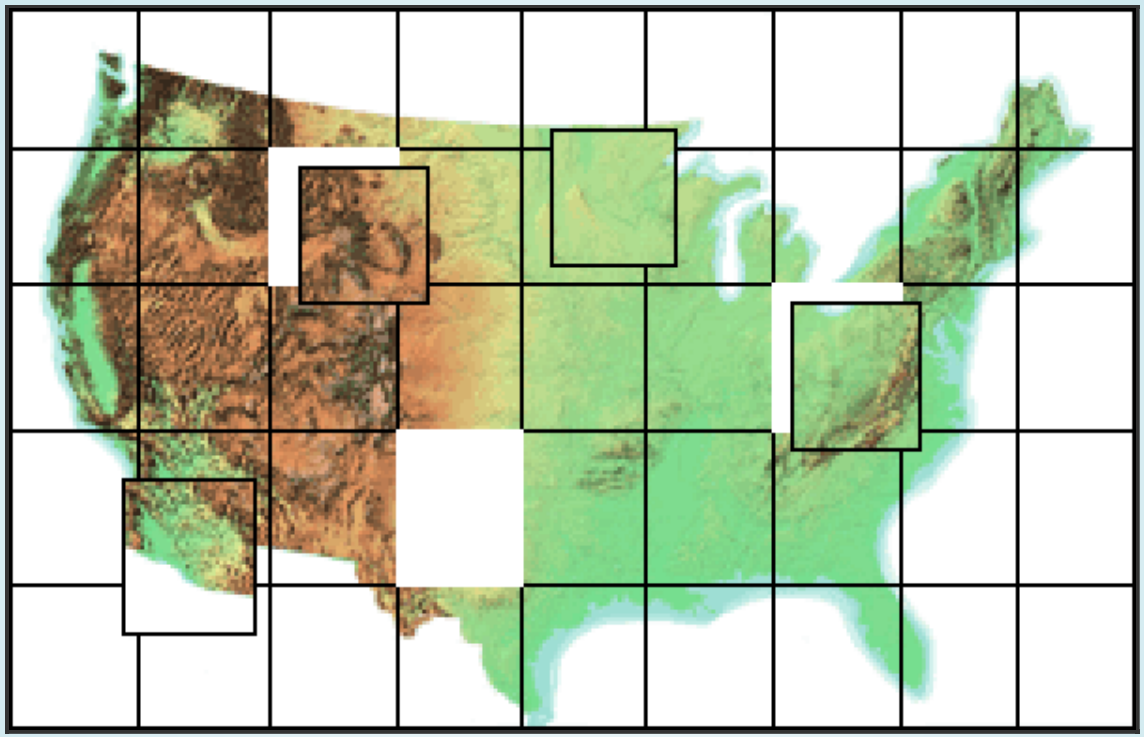


ACCUMULO

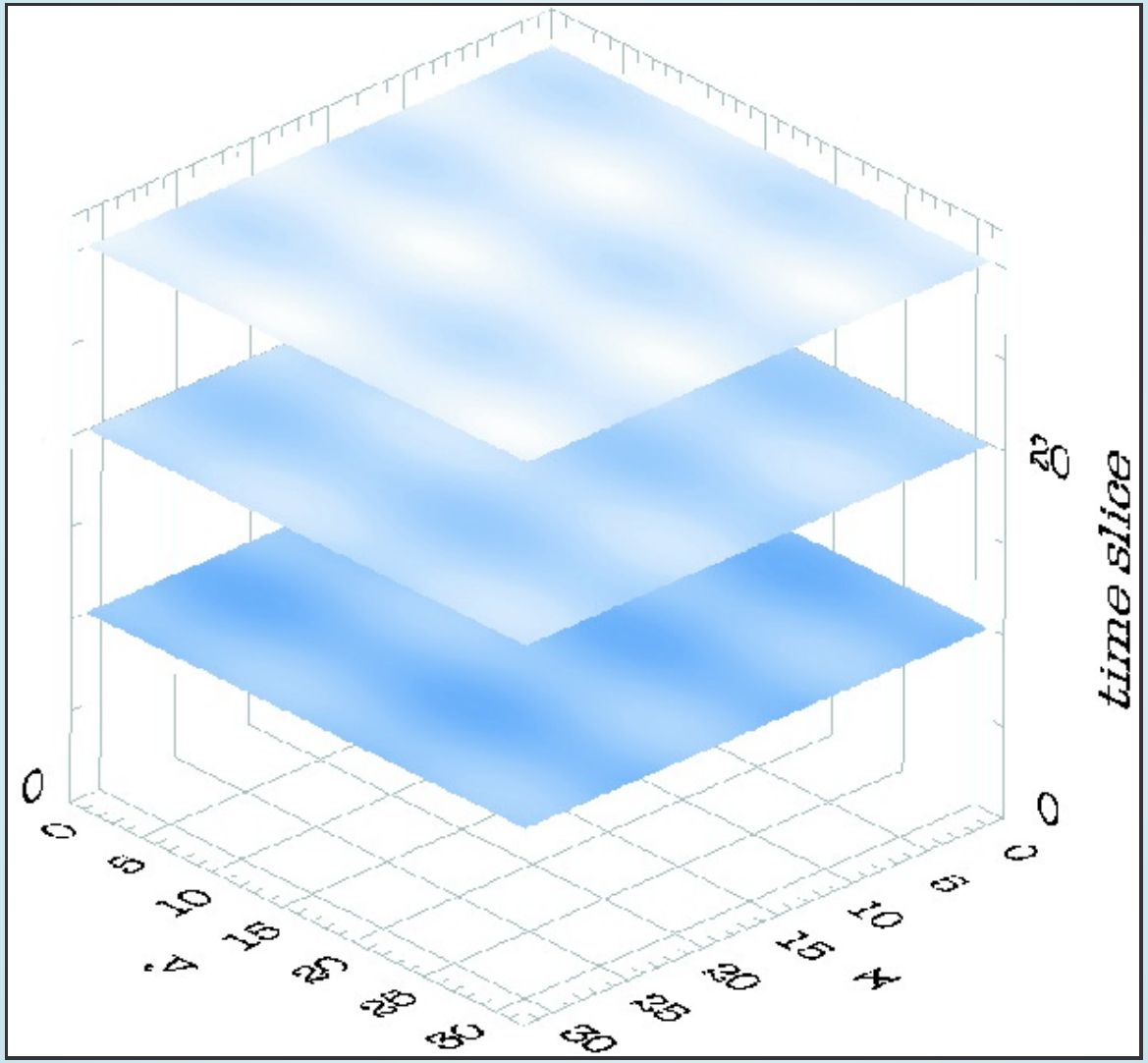
- Big table implementation
- Has sorted indexing
- Columnar database
- Also used by GeoMesa, another Scala project at LocationTech

STRATEGIES FOR WORKING WITH **BIG** RASTERS

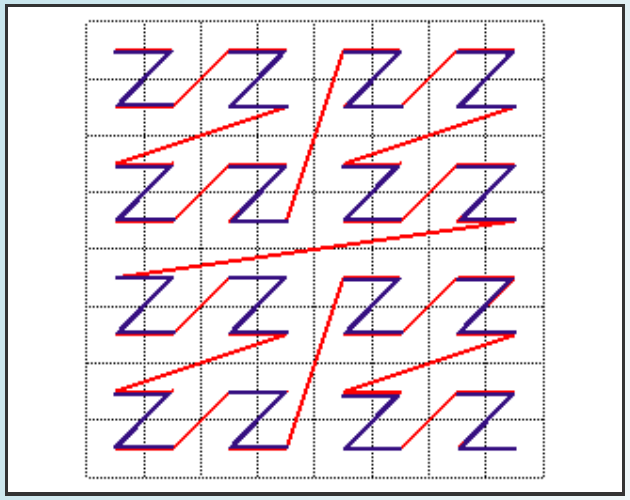
TILES

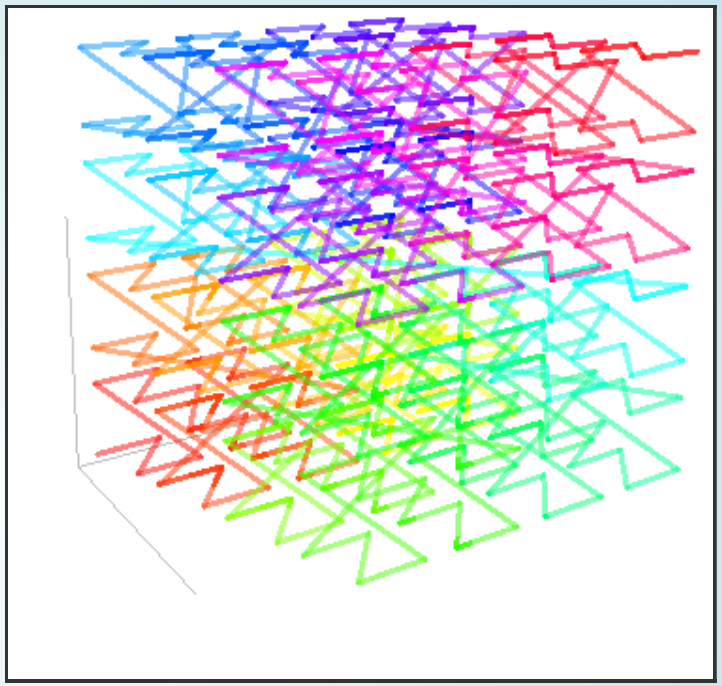


TILES



INDEXING TILES





RasterRDD[K]

K is key type, based on tile indexing.

- SpatialKey
- TemporalKey
- SpaceTimeKey

DATA LOADING

STEP 1:

EXPORT THE NETCDF DATA INTO 512X512 GEOTIFF TILES.

- Python code using GDAL and rasterio.
- AWS Auto scaling groups and SQS.
- Code: https://github.com/lossyrob/nex-chunker-worker

STEP 2:

INGEST THE DATA INTO ACCUMULO USING GEOTRELLIS-SPARK.

- Ingest the GeoTiffs to Accumulo in parallel across a cluster.
- Ingest consists of
  - reprojection
  - mosaicing to tile scheme (TMS)
  - pyramiding up zoom levels
  - Calculate index splits.

ANALYSIS OF NEX DATA

Live coding session...

THANKS!

Take it away Johan...

THE GEOTIFF FILE FORMAT

WITH

GEOTRELLIS AND SCALA

Johan Stenberg / @johanstenbergg

HOW DO YOU READ GEOTIFFS ON THE JVM?

- GDAL, GEOSPATIAL C LIB, FAST!
- GEOTOOLS, GEOSPATIAL JAVA LIB, SPEED?

WHY YET ANOTHER GEOTIFF READER?

- GeoTools large dependency
- GDAL Java bindings hard to install
- Go-To raster file format at GeoTrellis
- GeoTrellis is all about speed, everything optimized and benchmarked

WHAT IS THE GEOTIFF FILE FORMAT?

- Extension to the Tiff File Format
- Used for images with Geospatial Metadata
- Adds a bounding box and the CRS through tags

GEODATA?

- Bounding Box easy to read
- Coordinate Reference System horrible to read
- Turn it into a proj4 string and use the proj4j lib to read

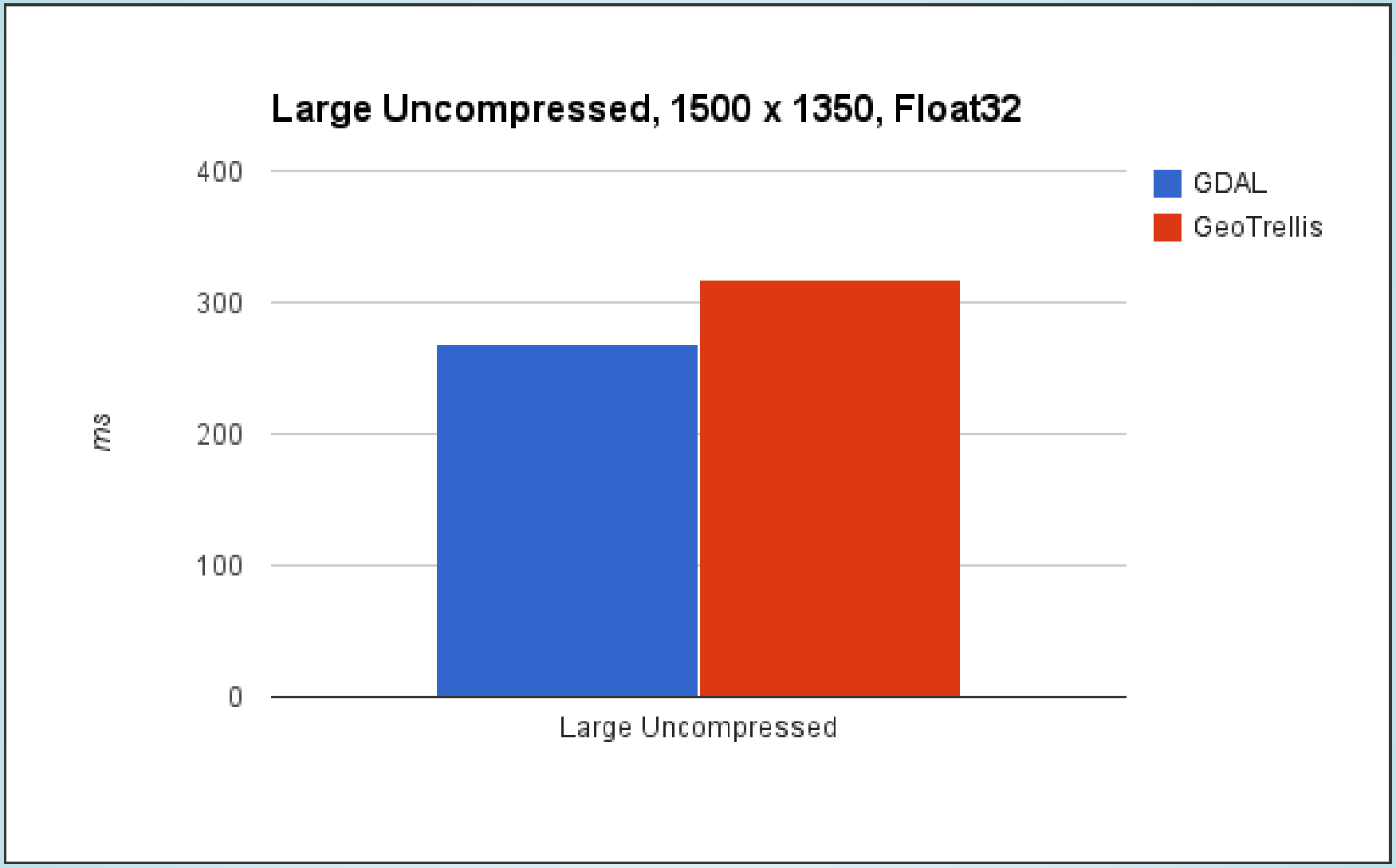
COMPRESSIONS

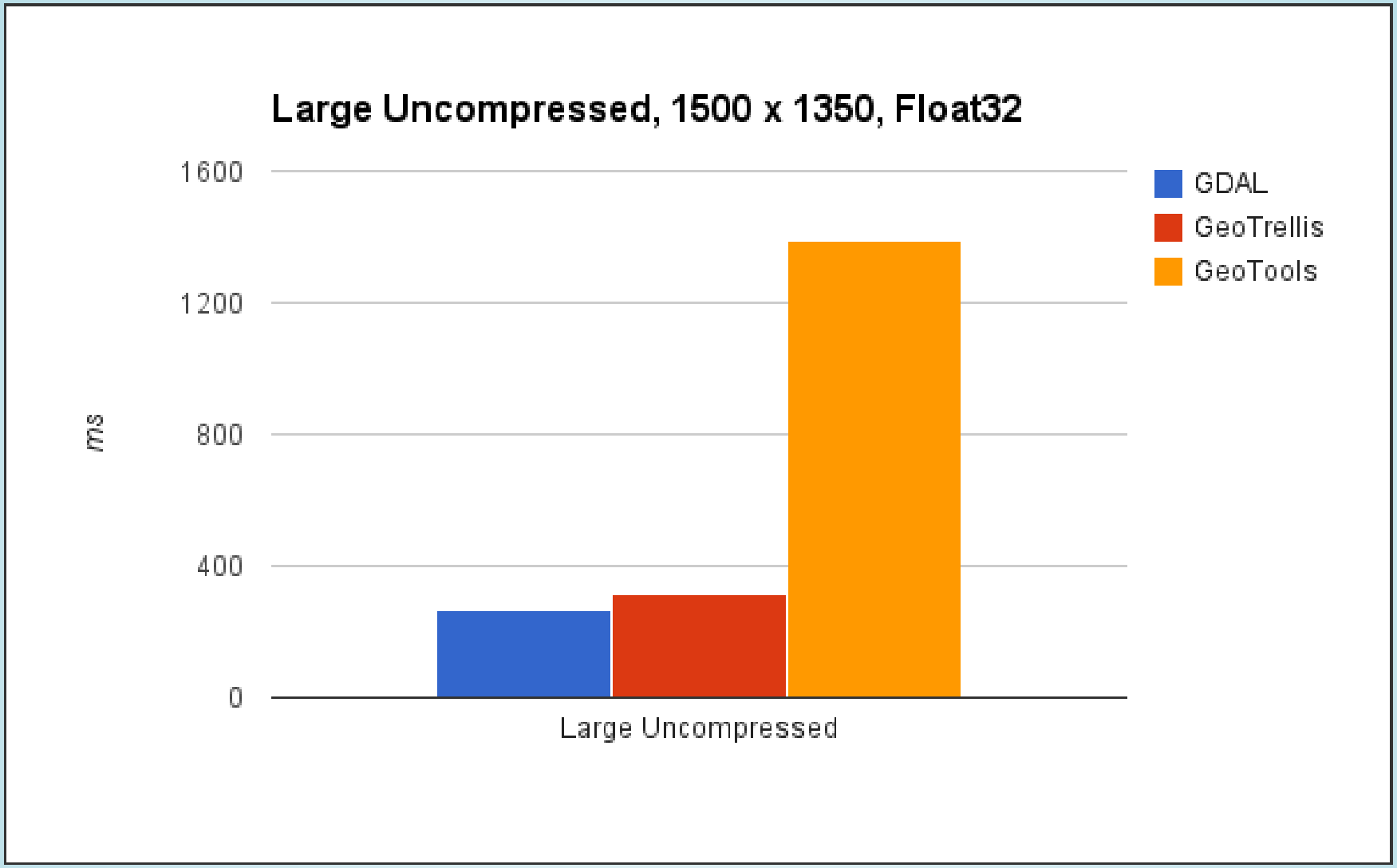
- Huffman, CCITT3, CCITT4, Packbits
- LZW
- Zip

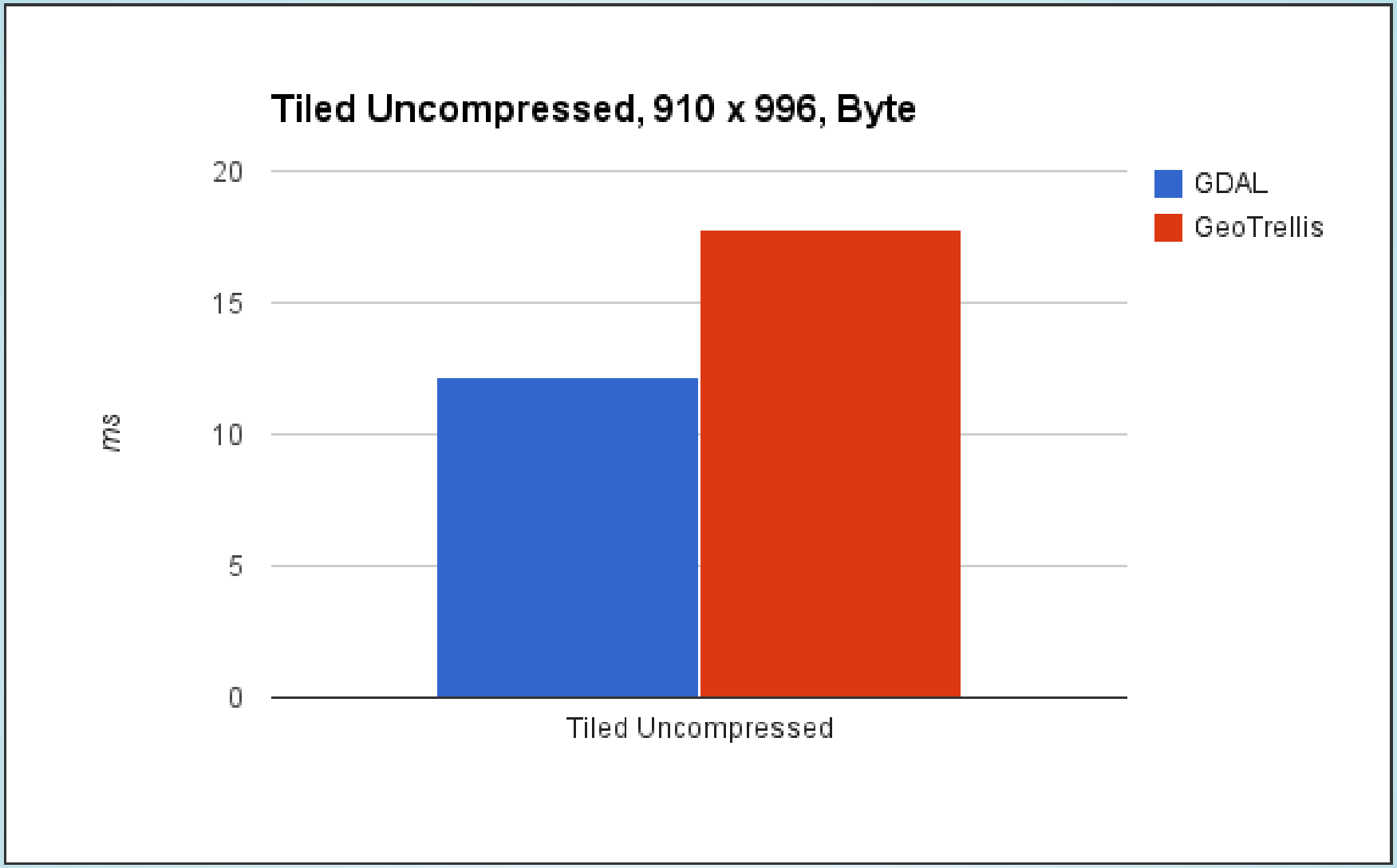
BENCHMARK TIME!

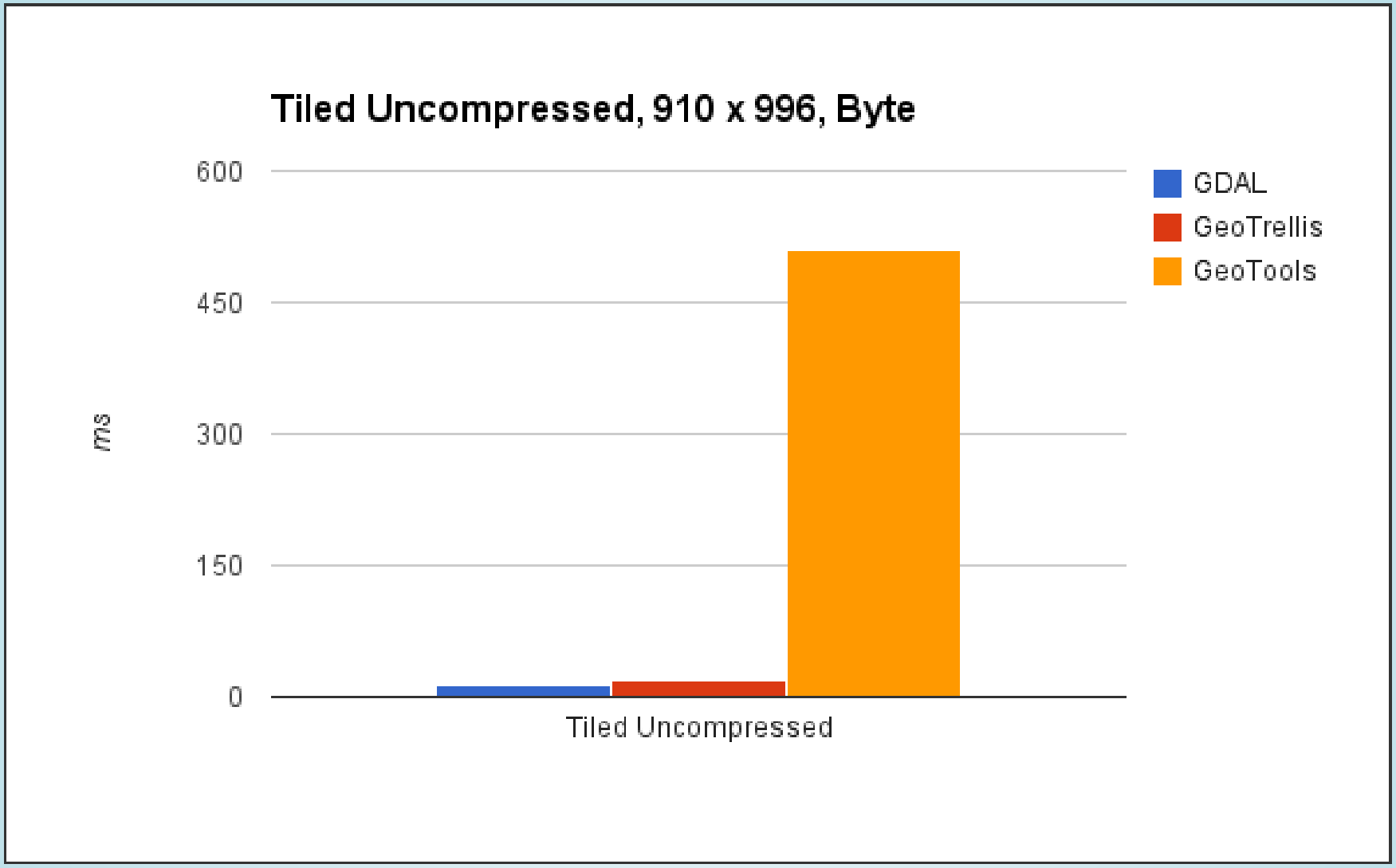
BENCHMARK DISCLAIMER

- Ran on my development computer
- Conducted with Caliper
- Microbenchmarks, look at relative speed, not speed
- GDAL is read through the Java bindings, into GeoTrellis rasters
- GeoTools is also turned into GeoTrellis rasters

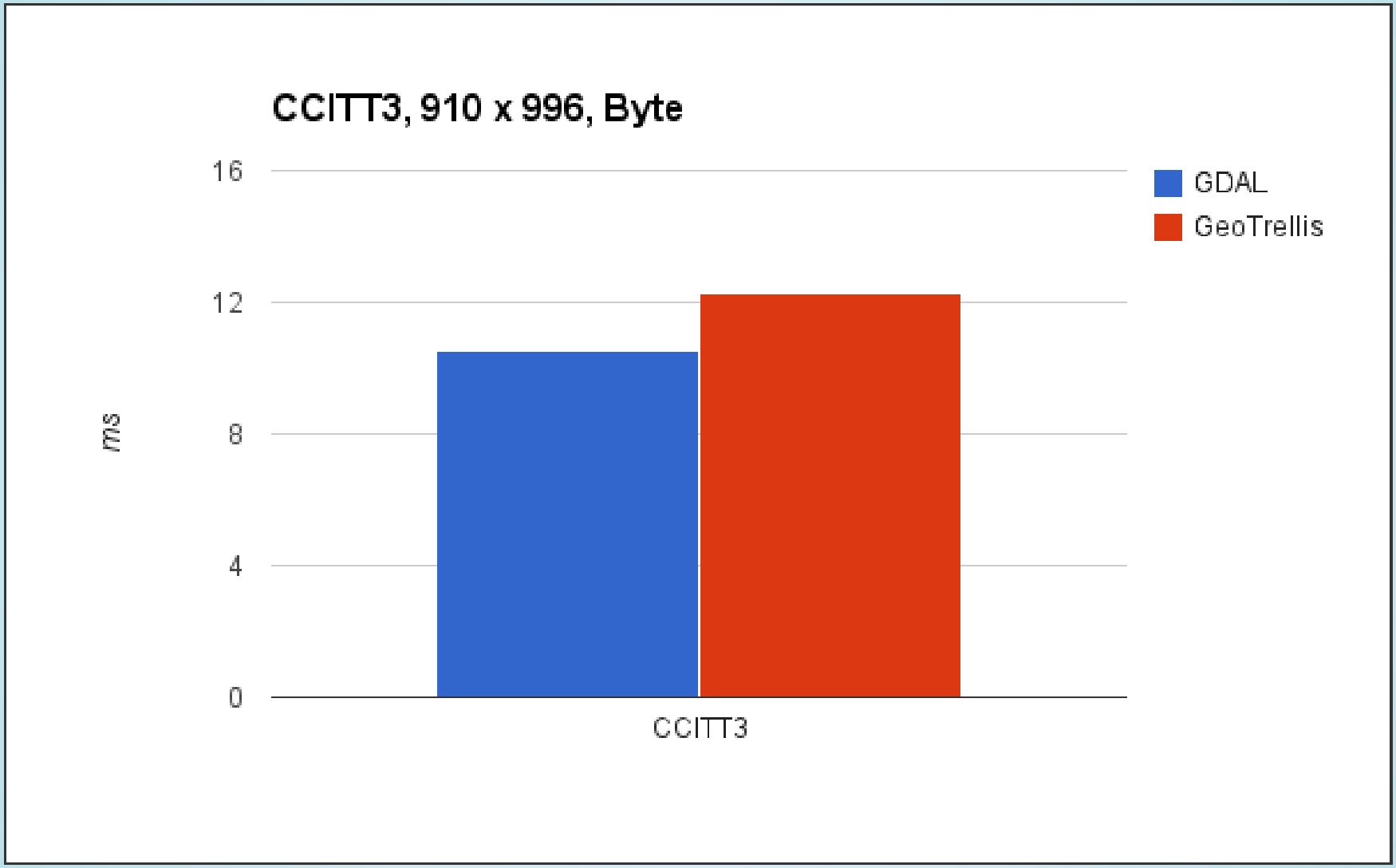

Large Uncompressed, 1500 x 1350, Float32
400
300
200
100
0
ms
GDAL
GeoTrellis
Large Uncompressed


Large Uncompressed, 1500 x 1350, Float32
1600
1200
800
400
0
ms
Large Uncompressed
GDAL
GeoTrellis
GeoTools

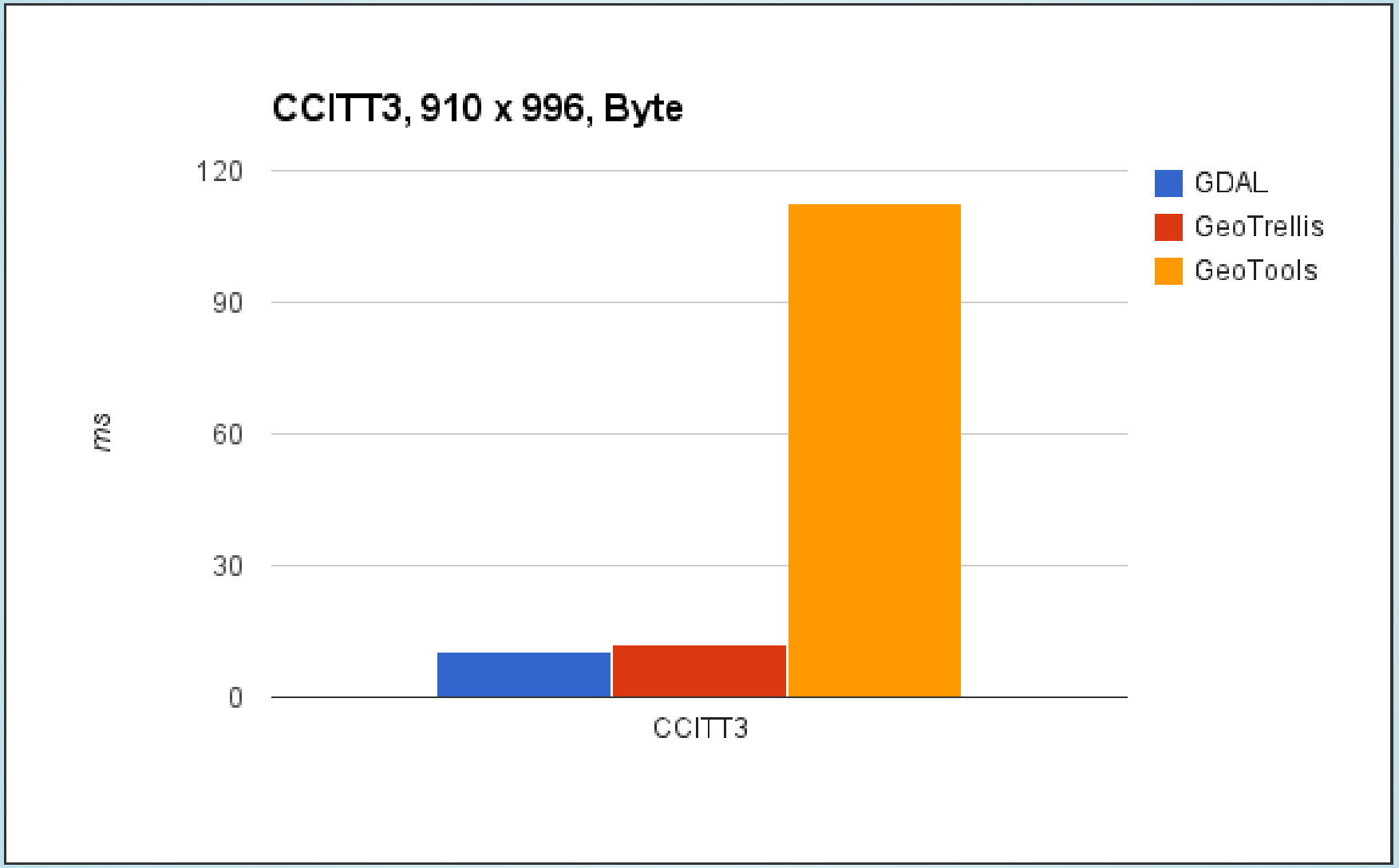

Tiled Uncompressed, 910 x 996, Byte
20
15
10
5
0
ms
Tiled Uncompressed
GDAL
GeoTrellis

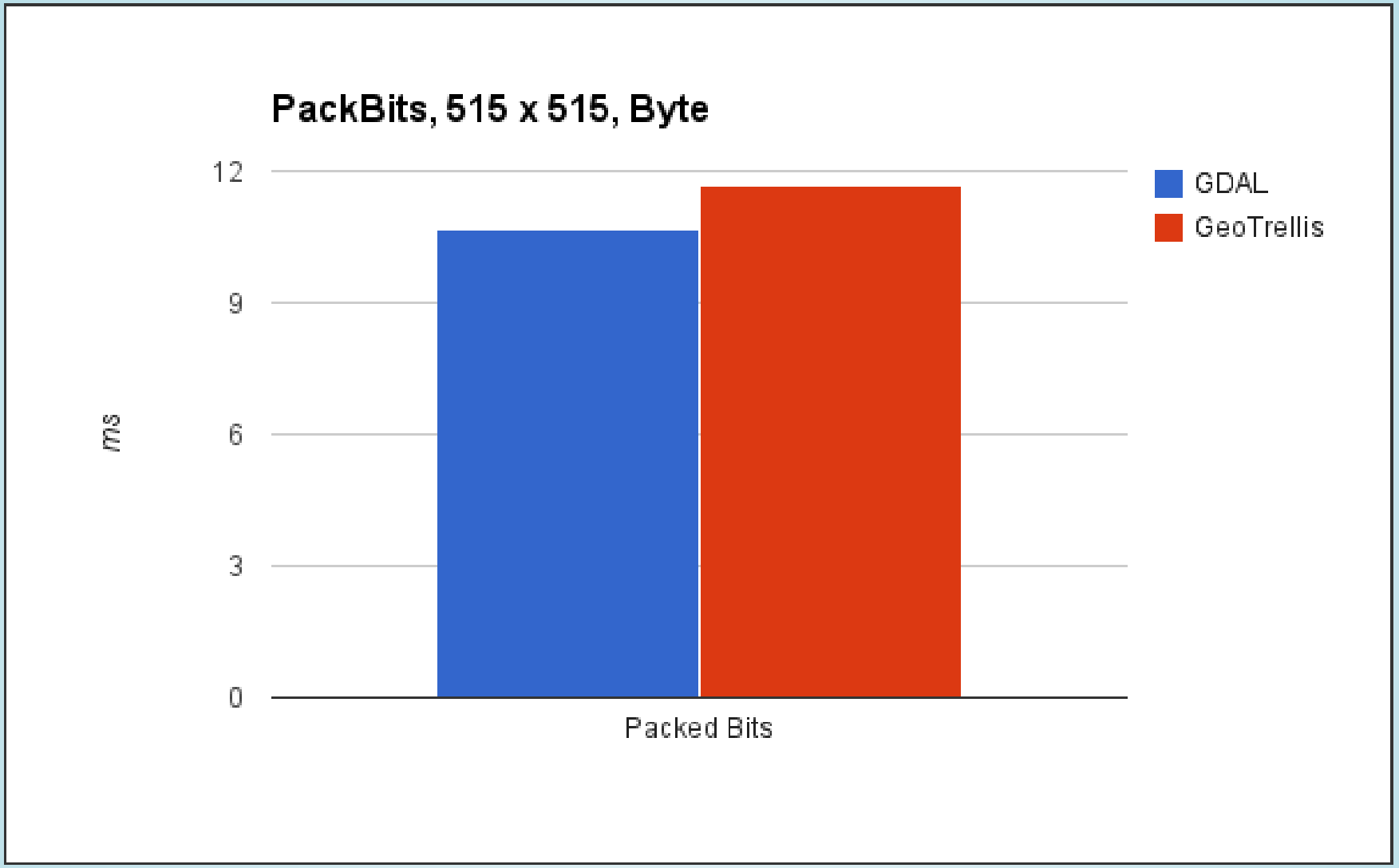

Tiled Uncompressed, 910 x 996, Byte
600
450
300
150
0
ms
Tiled Uncompressed
GDAL
GeoTrellis
GeoTools

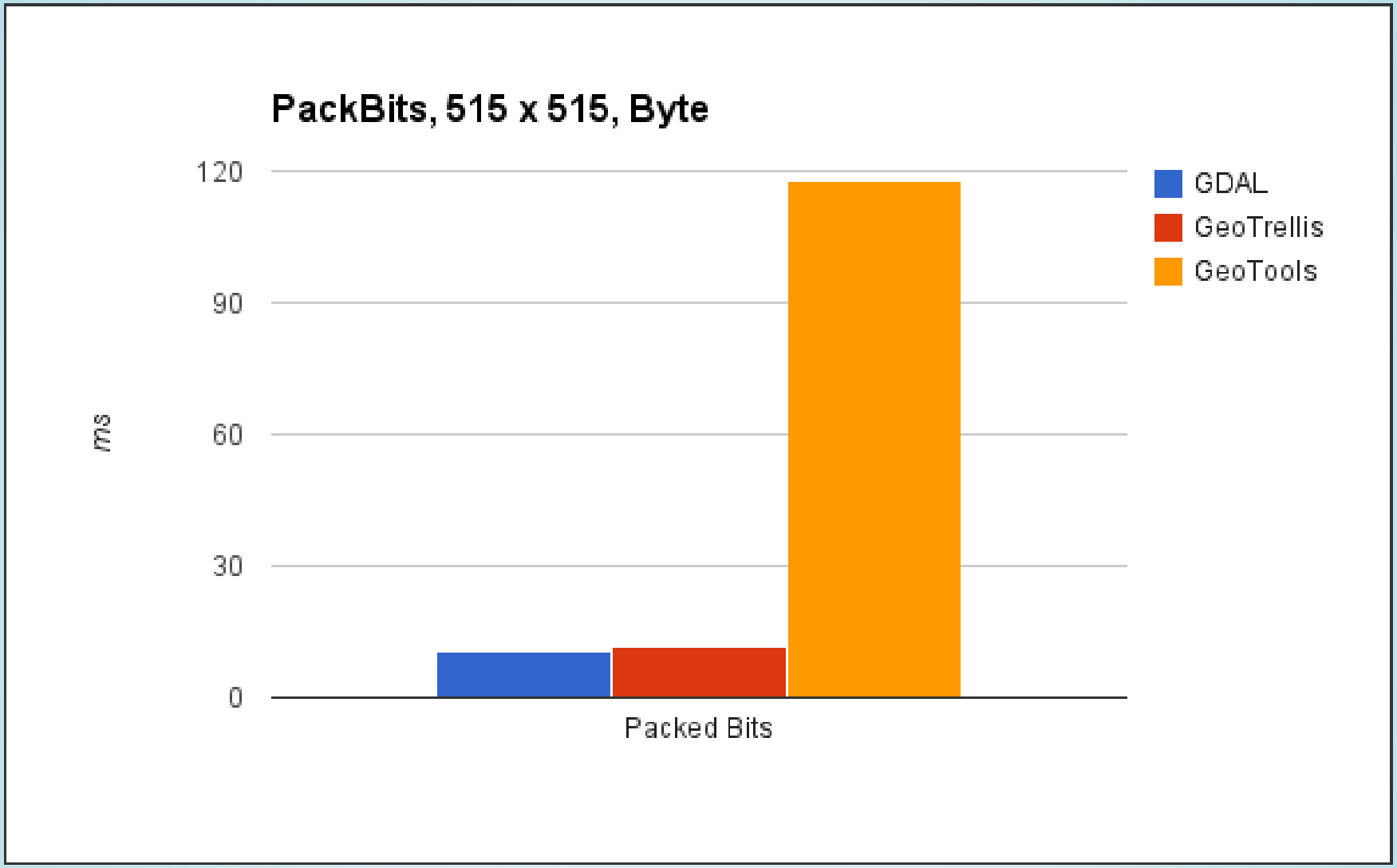
~SAME FOR CCITT3 AND CCITT4

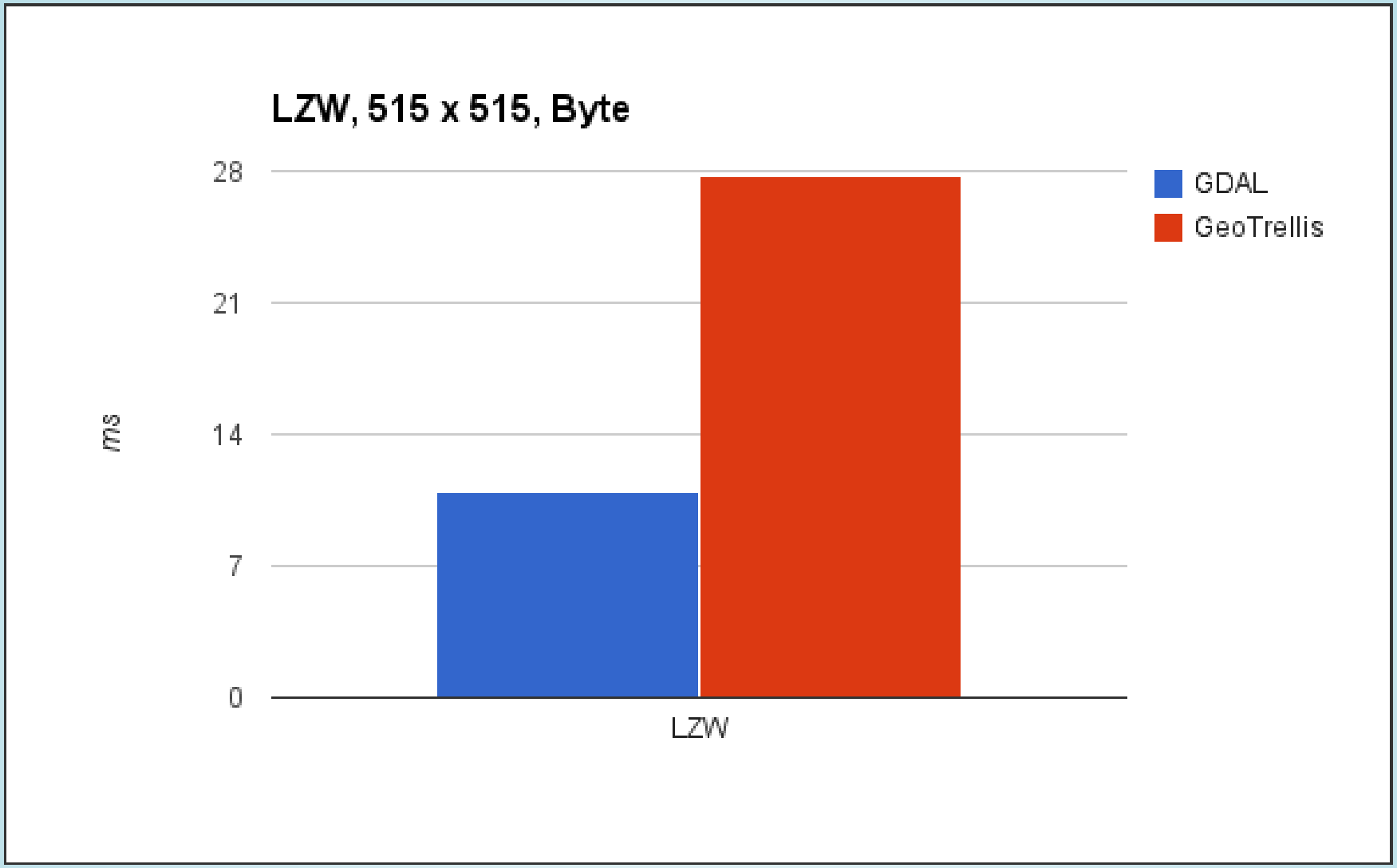


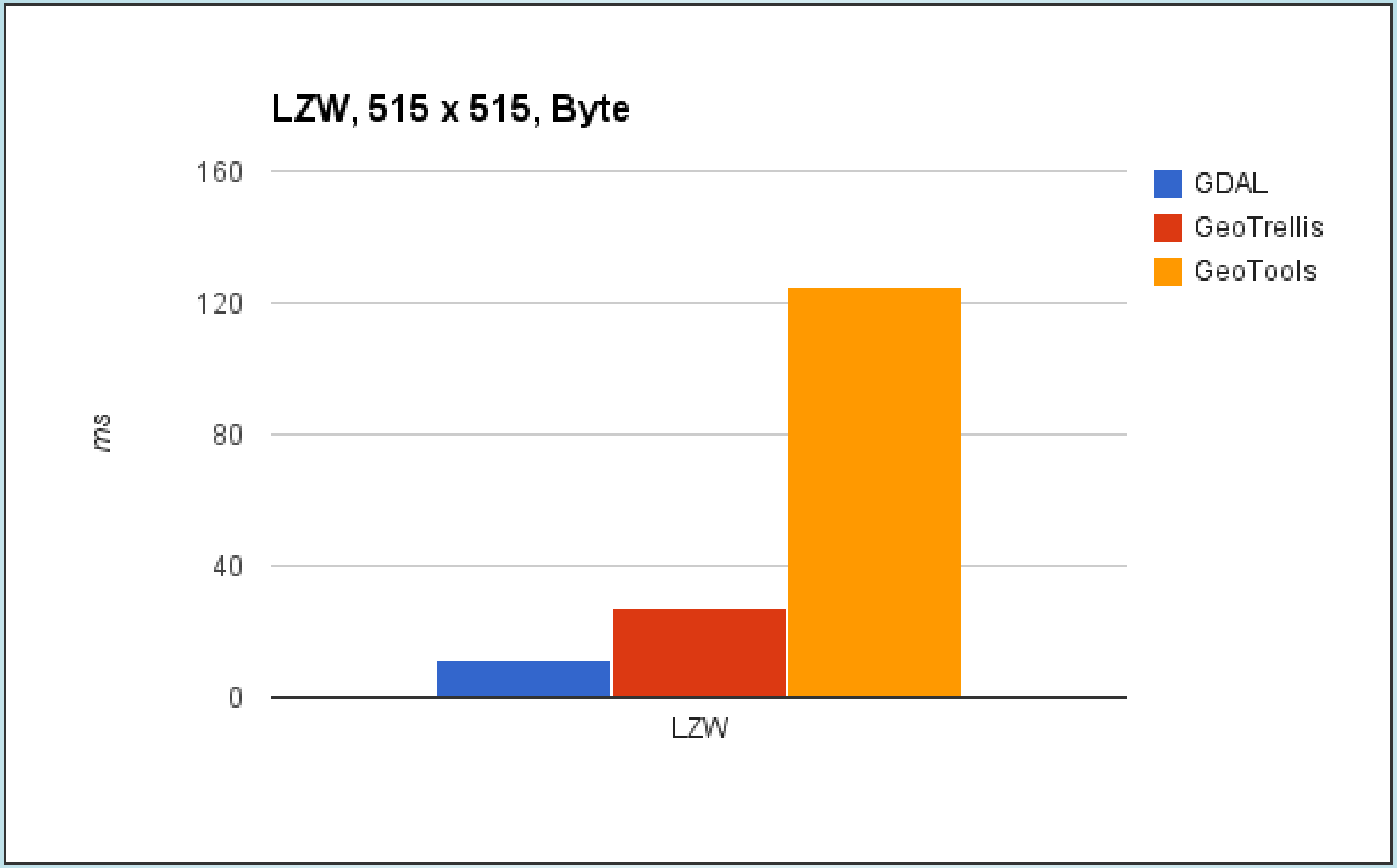
~SAME FOR CCITT3 AND CCITT4




PackBits, 515 x 515, Byte
12
9
6
3
0
ms
Packed Bits
GDAL
GeoTrellis


PackBits, 515 x 515, Byte
120
90
60
30
0
ms
Packed Bits
GDAL
GeoTrellis
GeoTools


LZW, 515 x 515, Byte
28
21
14
7
0
ms
LZW
GDAL
GeoTrellis


LZW, 515 x 515, Byte
160
120
80
40
0
ms
LZW
GDAL
GeoTrellis
GeoTools