Support for massive spatial datasets in GRASS GIS

Markus Metz

Prague FCE CTU 2011

Massive spatial datasets

Large is relative to

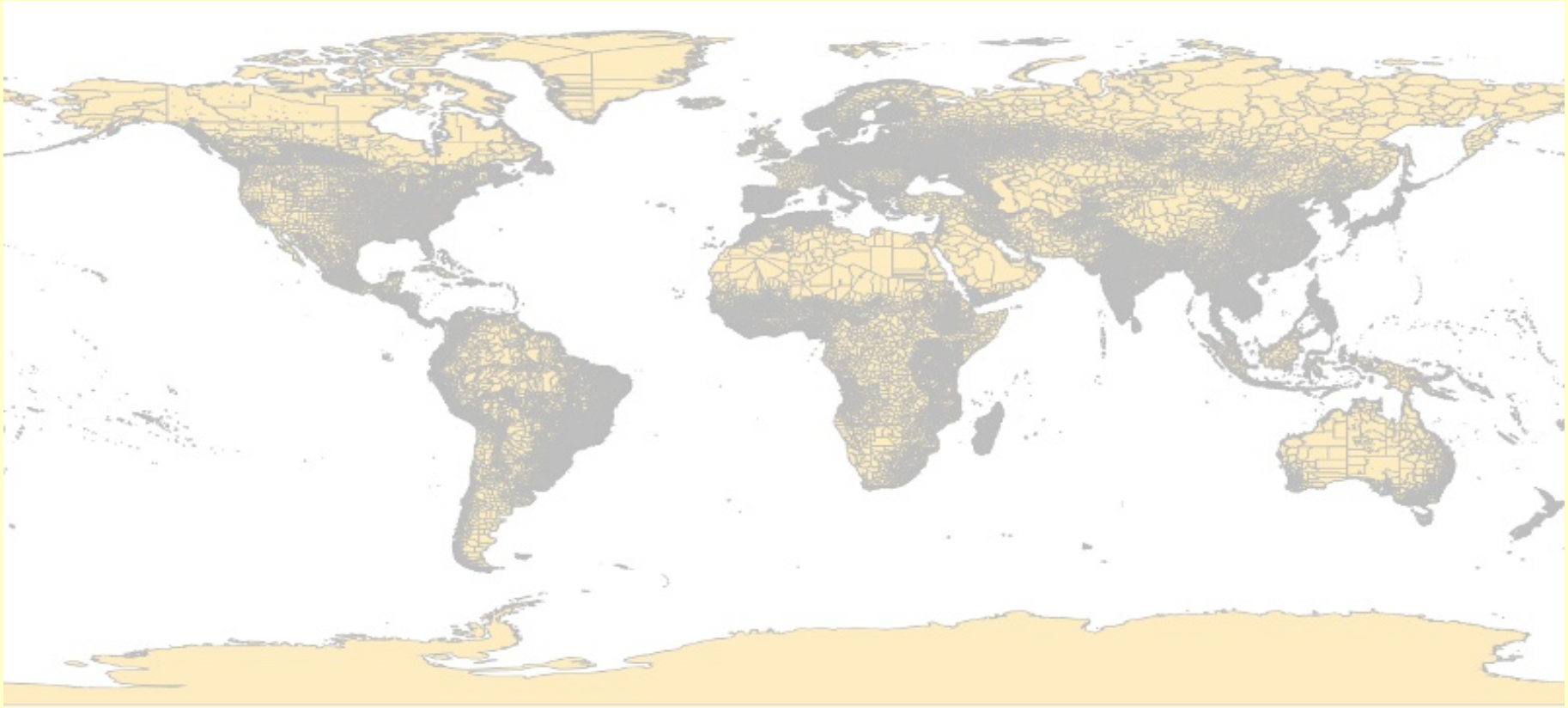
- RAM
- disk space
- processing time
- largest supported file size

Massive spatial datasets

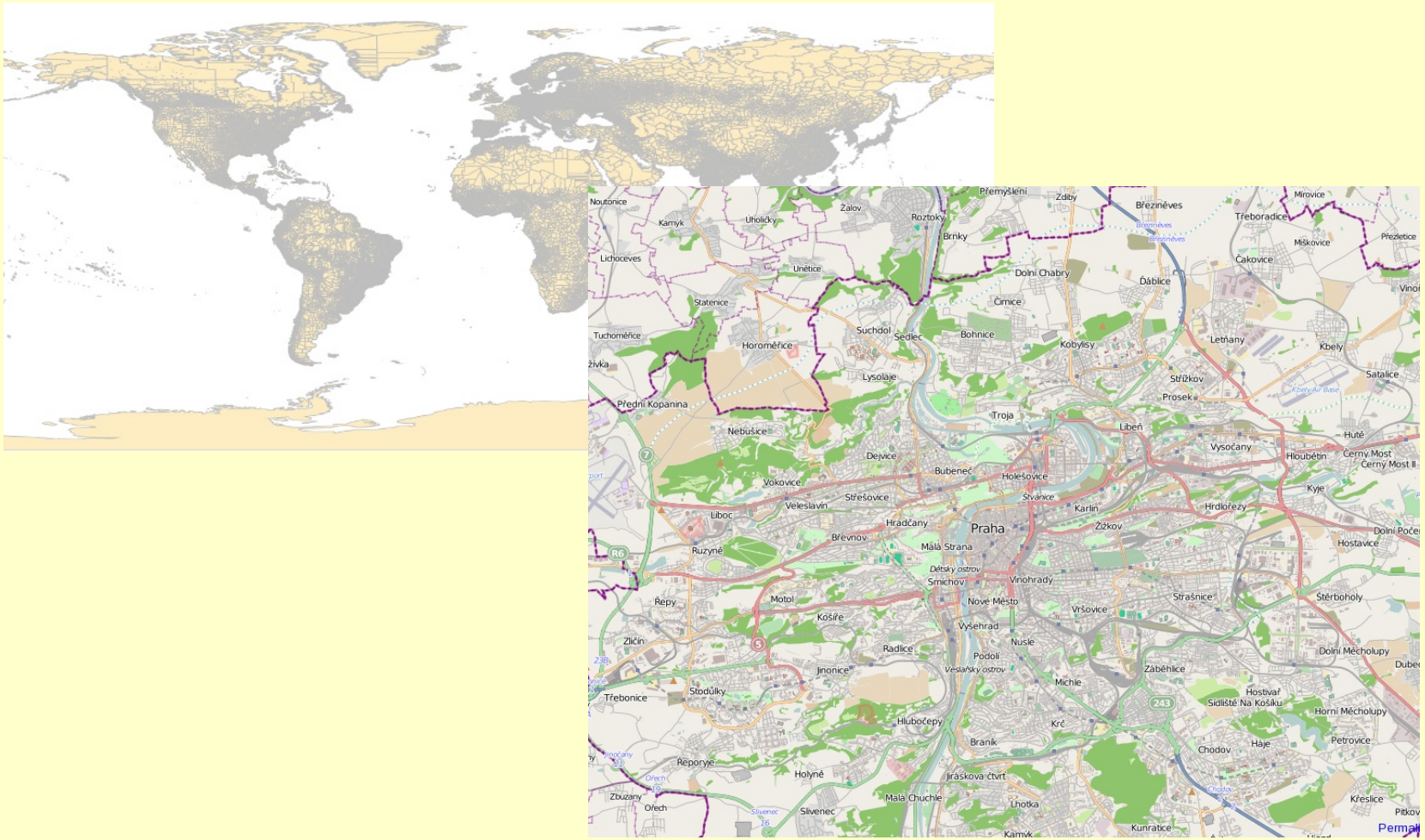
Examples

- Global Administrative Areas (1.3 GB)
- OpenStreetMap (20 GB)
- Corine Land Cover (vector version, 4.4 GB)
- LiDAR (open ended size)
- DEMs (open ended size)

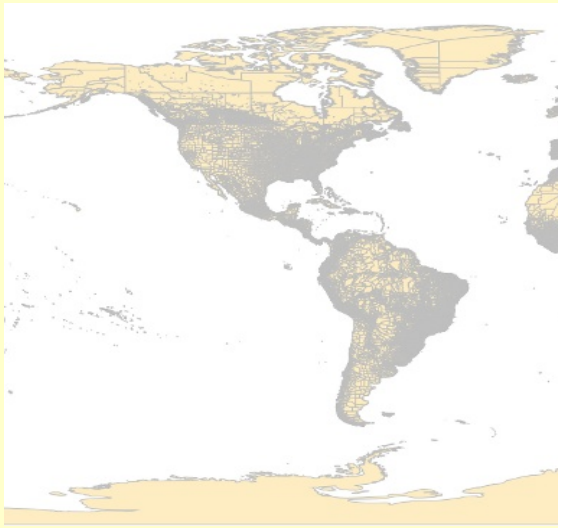
Massive spatial datasets

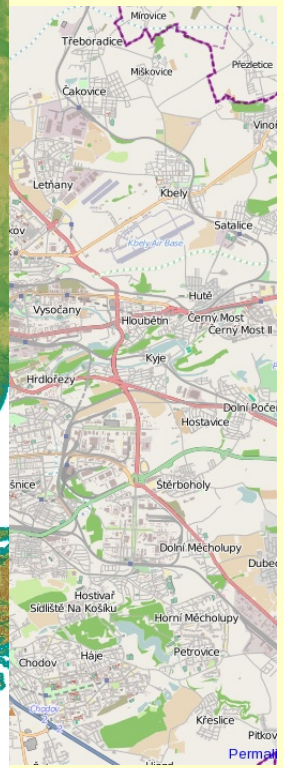


Massive spatial datasets



Massive spatial datasets





Massive spatial datasets

Particularly problematic for

- Least cost searches
- Hydrological modelling
- Image reprojection and (ortho-)rectification
- Vector topology

Data processing concepts

Raster

Only few processes need to load the full raster at once

- Scattered (random) access
- Sweep-lines
- Sorting and searching

Vector

loading the full vector map to memory is probably never necessary

BUT: processing time can be long,
support datastructures need to be loaded

Data processing concepts

Raster

Scattered (random) access -> never happens

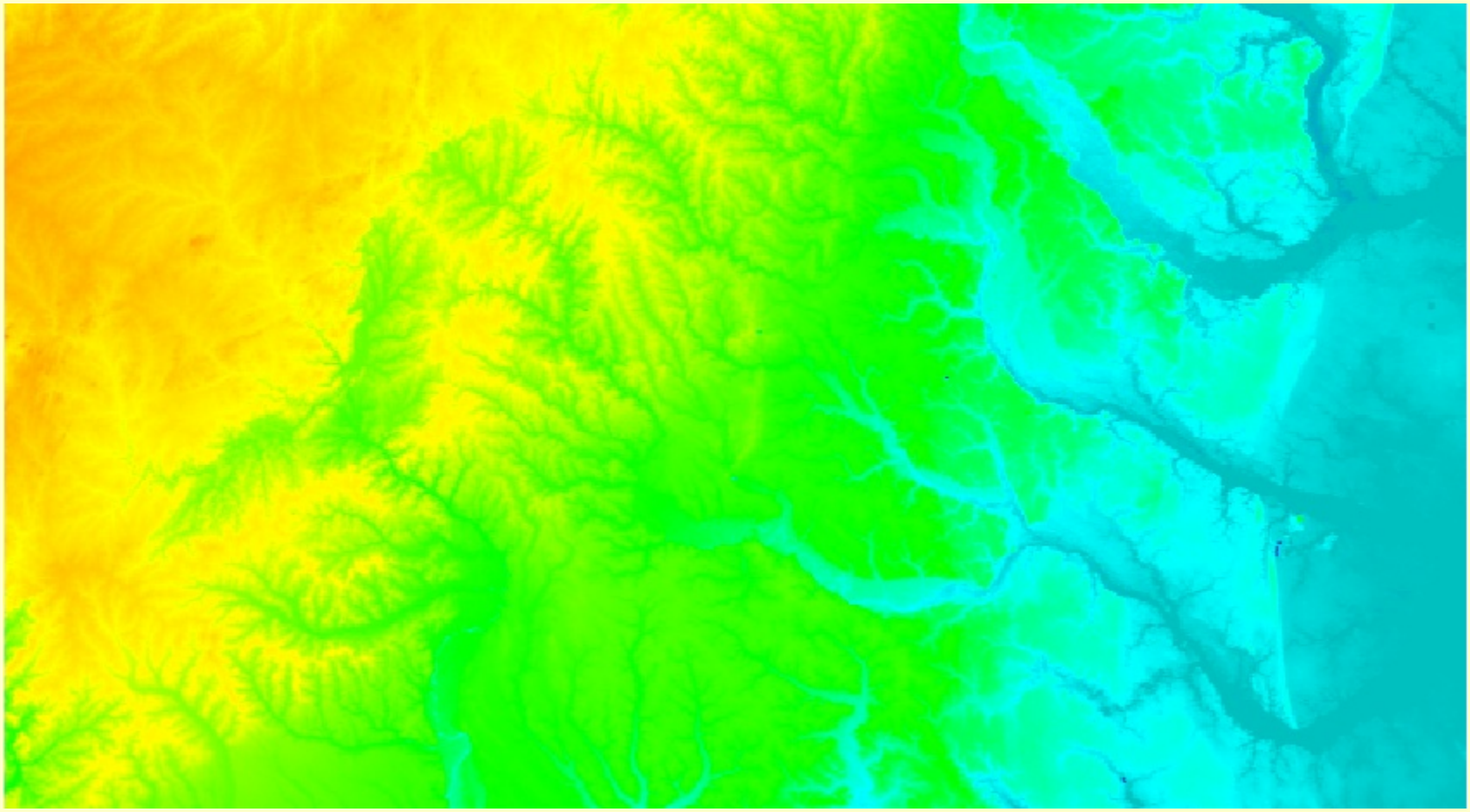
Data processing concepts

Raster

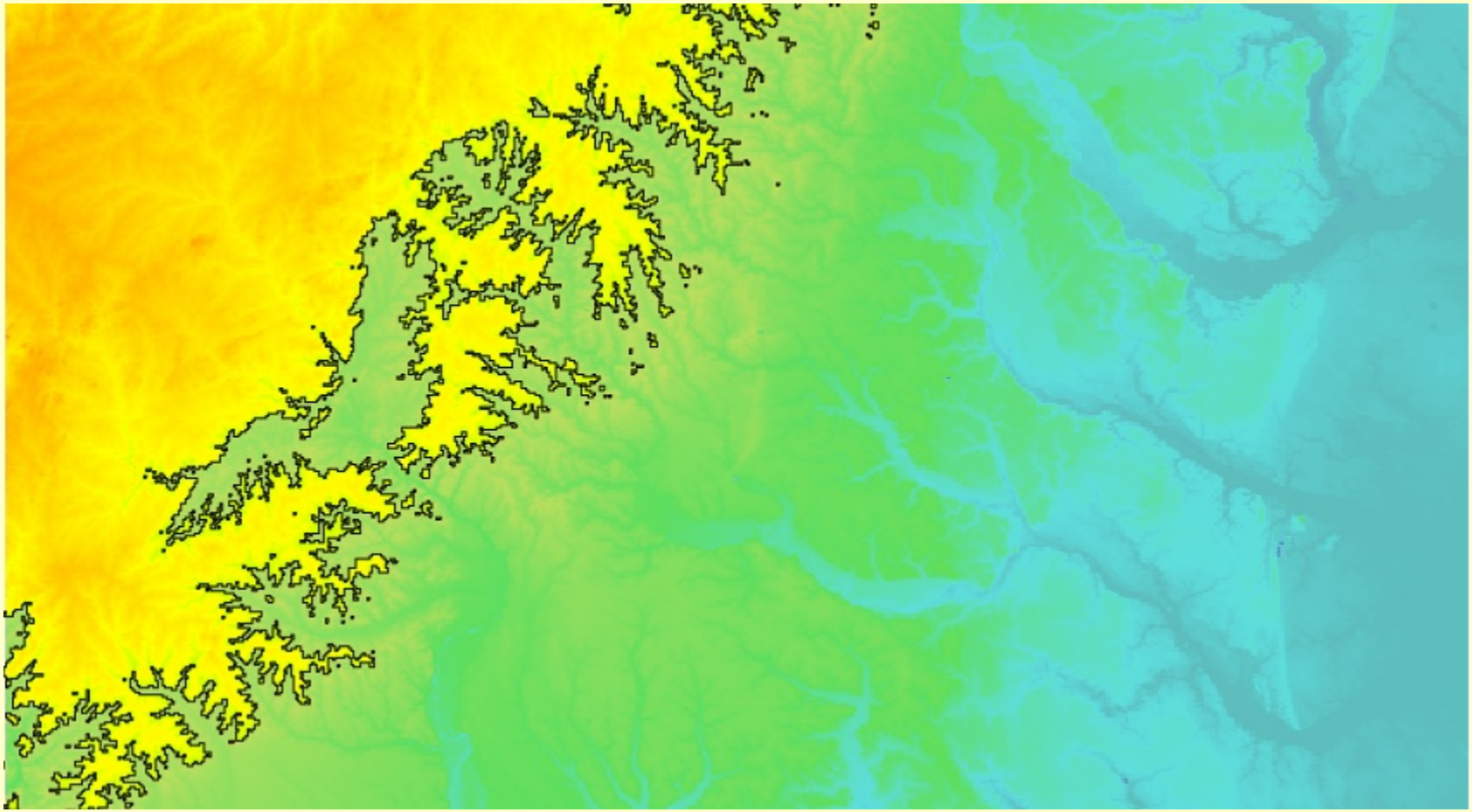
Scattered (random) access -> never happens

Sweep-lines: e.g. hydrology, cost surfaces

Raster sweep line



Raster sweep line



Data processing concepts

Raster

Scattered (random) access -> never happens

Sweep-lines: e.g. hydrology, cost surfaces

→ Sorting and searching

Massive spatial datasets

Solutions

- **NEVER** load the full raster to memory
- Tiling
- external memory
- Fast sorting and searching
- Large File Support (LFS), files > 2 GB

... available in GRASS GIS

What is GRASS GIS ?

Geographic **R**esources **A**nalysis **S**upport **S**ystem

Free and Open Source since 1984

Member of OSGeo Your Open Source Compass

Linked to GDAL, QGIS and R

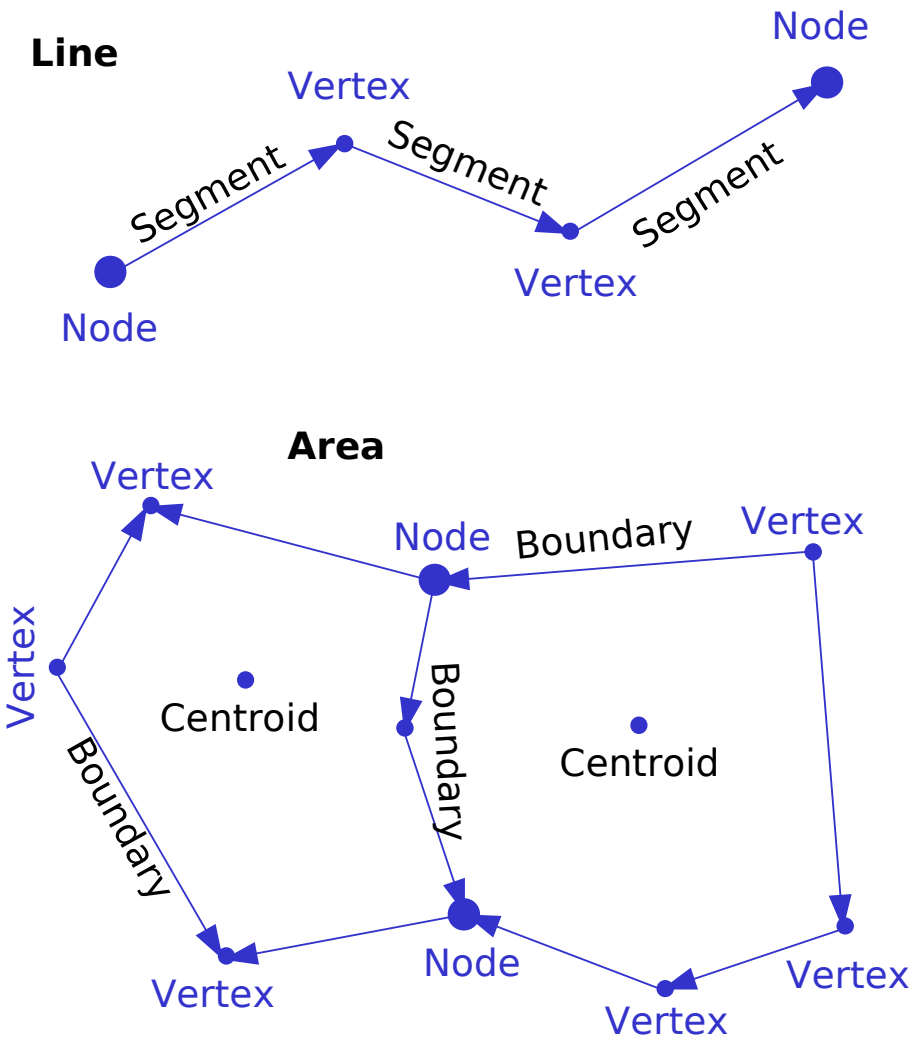
Portable: GNU/Linux, Mac OSX, MS-Windows, SUN, etc.

GRASS vector model

Vector geometry types

- Point
- Centroid
- Line
- Boundary
- Area (Boundaries + Centroid)
- Face (3D Area)
- [Kernel (3D Centroid)]
- [Volumes (Faces + Kernel)]

All types are **true 3D**: x, y, z



OGC Simple Features vs Vector Topology

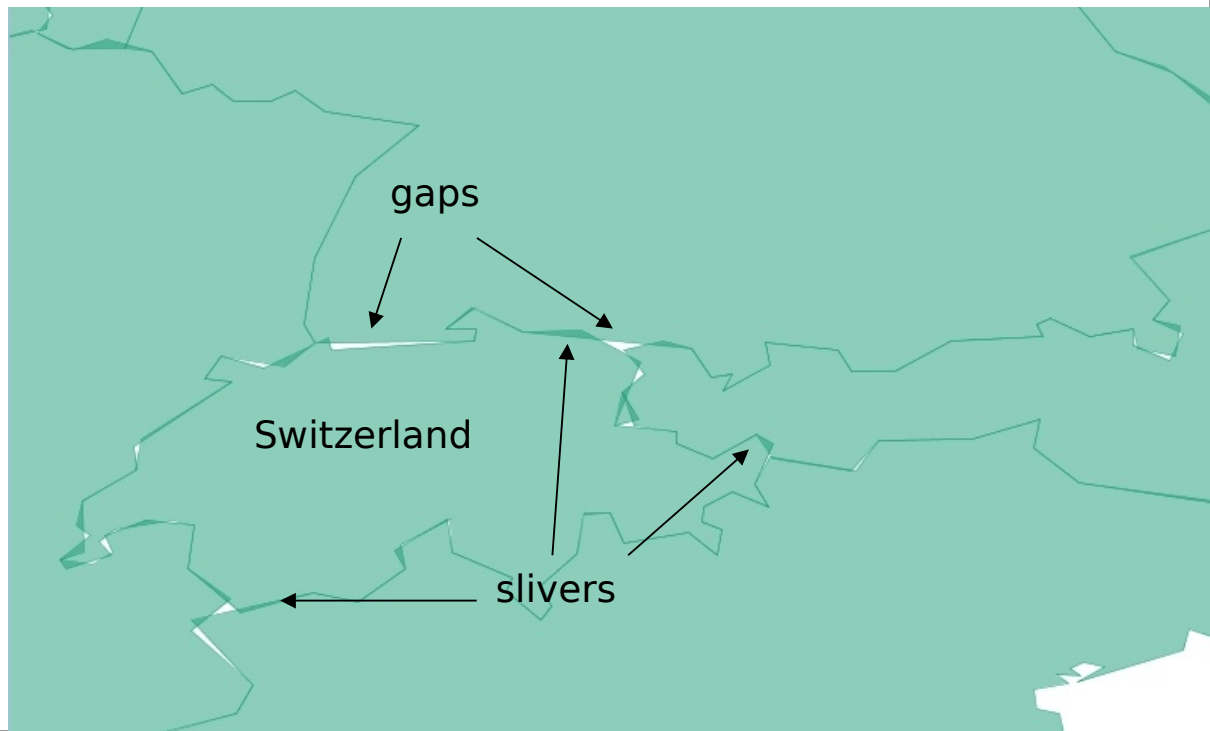
OGC Simple Features

points, lines, polygons

-> replicated boundaries for adjacent areas

faster computations, but extra work for maintenance

Non-topological
polygons generalized



OGC Simple Features vs Vector Topology

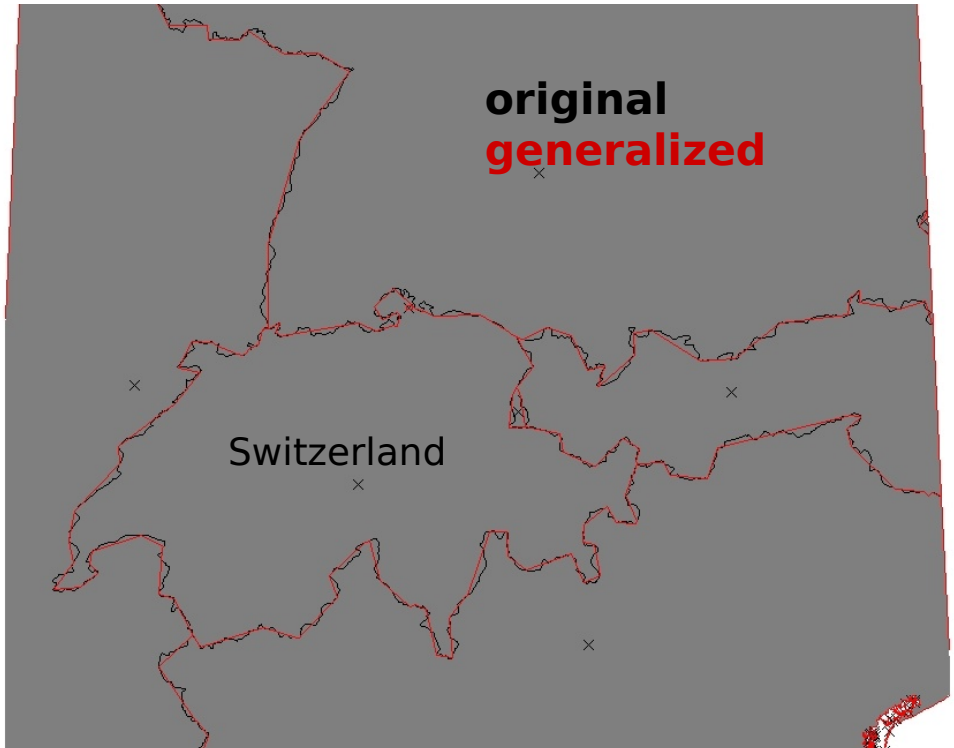
Vector Topology

areas are constructed from boundaries

boundaries are shared between adjacent areas

slower computations, but less (nearly no manual) maintenance

Topological
boundaries
generalized



Data processing in GRASS GIS

Raster

Developed in the 1980's

row by row	*default*
segmented	*reprojection, interpolation, hydrology, ...*
sorted	*terraflow, viewshed*

Vector

Mix of external memory and all in memory (topology)

- network analysis
- LiDAR point clouds

Massive datasets in GRASS 7

Improved raster tiling

Improved external memory, raster and vector

Fast sorting and searching

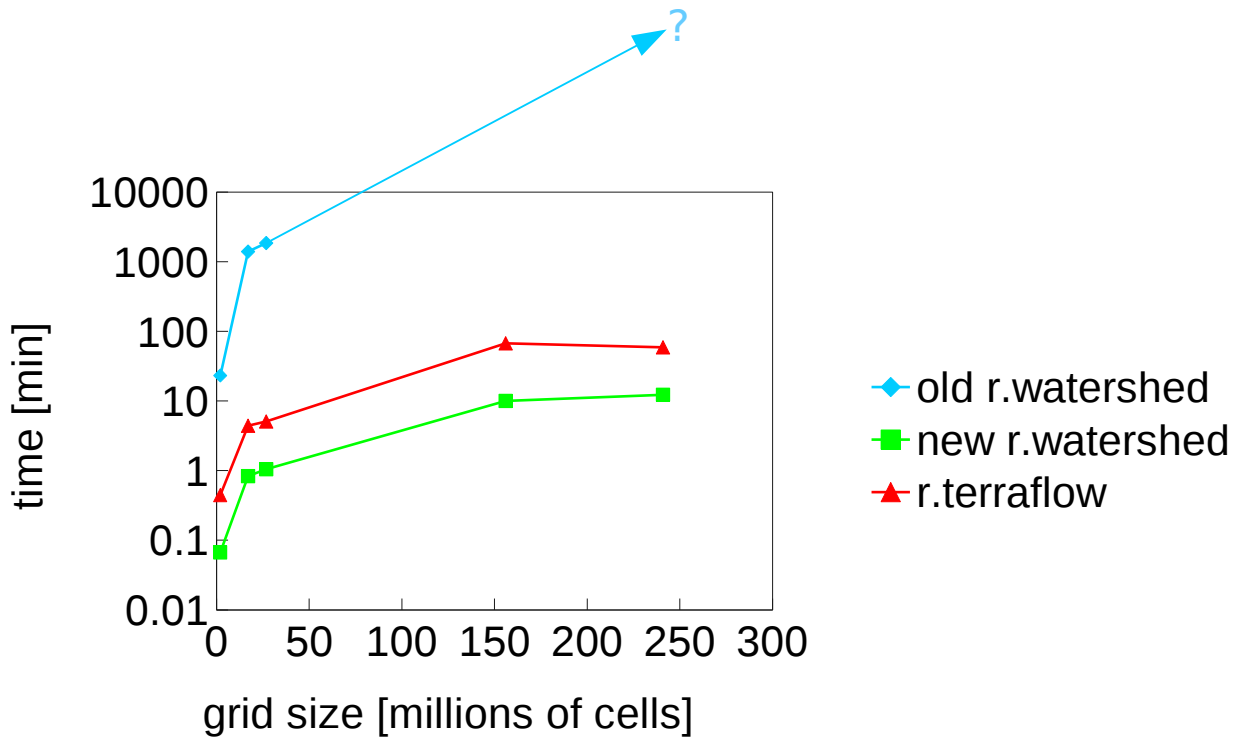
Large File Support (LFS), files > 2 GB

Massive datasets in GRASS 7

Raster processing

Massive datasets in GRASS 7

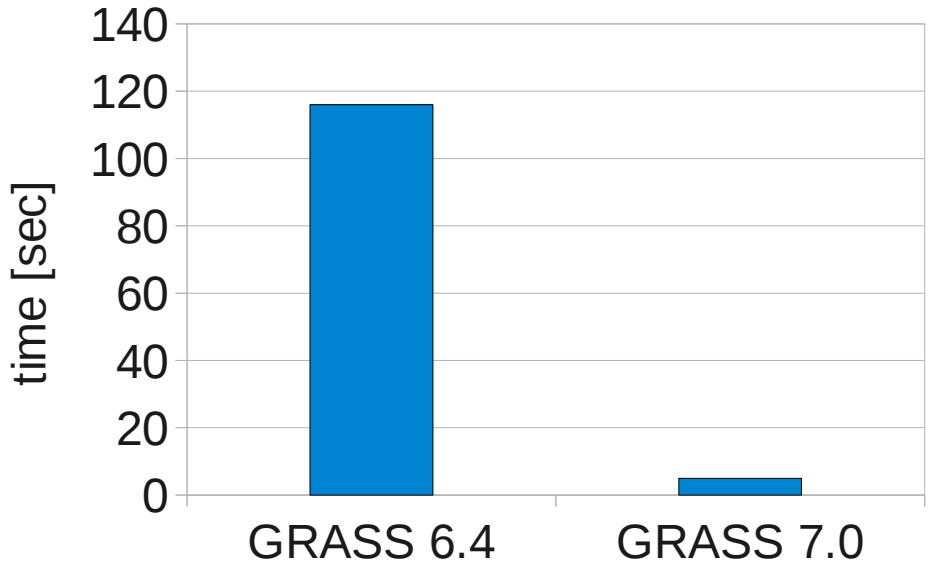
Hydrology: r.watershed (memory + processing time)



Massive datasets in GRASS 7

Cost surfaces: *r.cost*

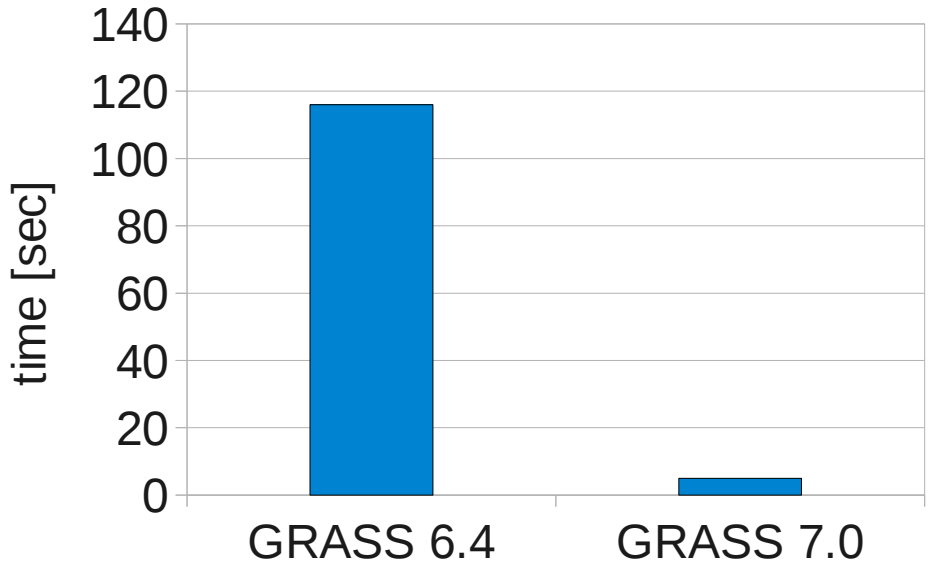
4.5 million cells, 160 start points



Massive datasets in GRASS 7

Cost surfaces: *r.cost*

4.5 million cells, 160 start points

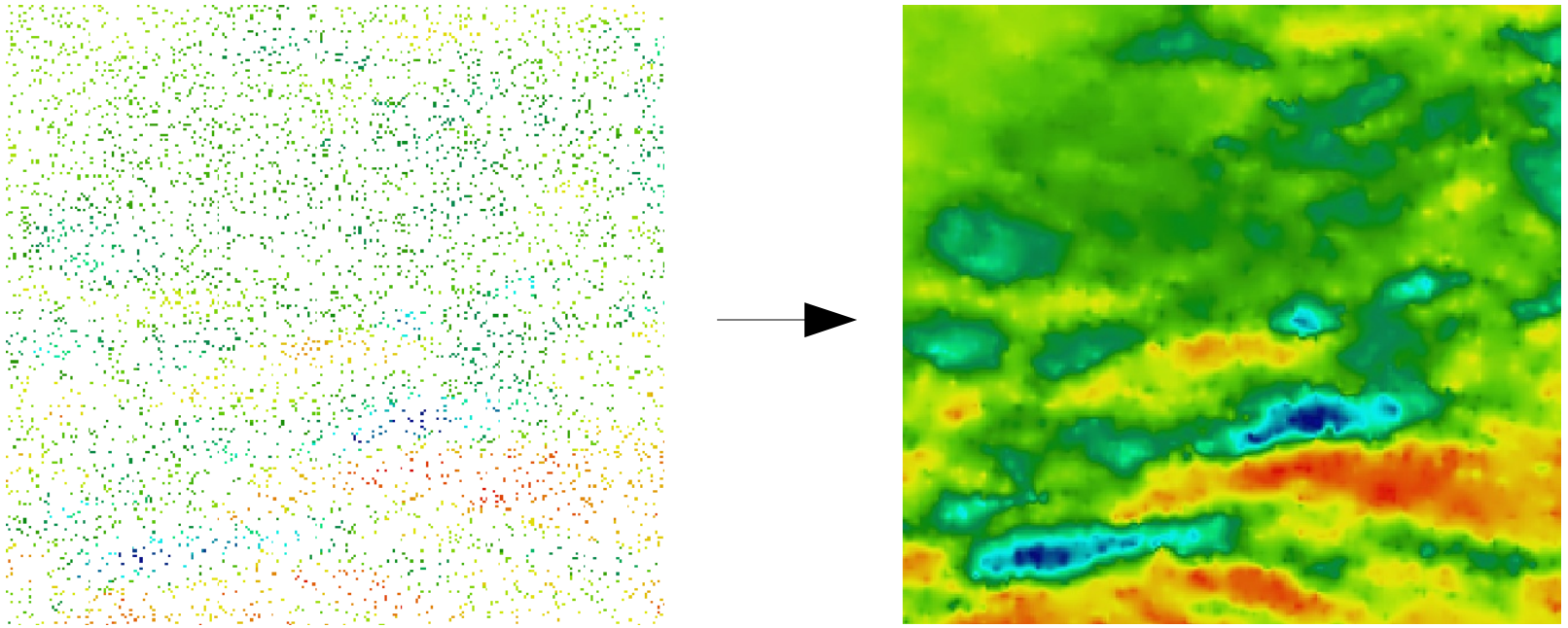


Similar for viewshed analysis: r.los ↔ r.viewshed

Massive datasets in GRASS 7

B-spline raster interpolation

Raster or vector points as input

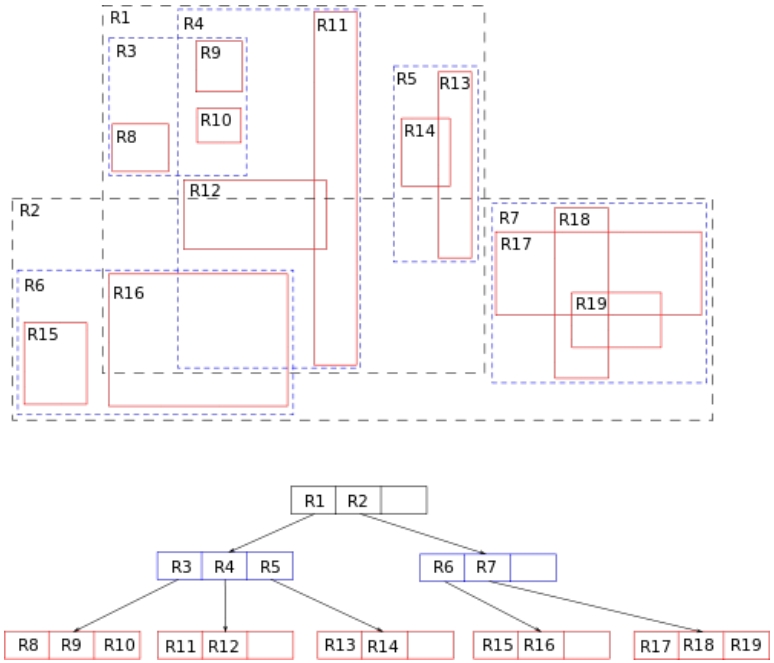


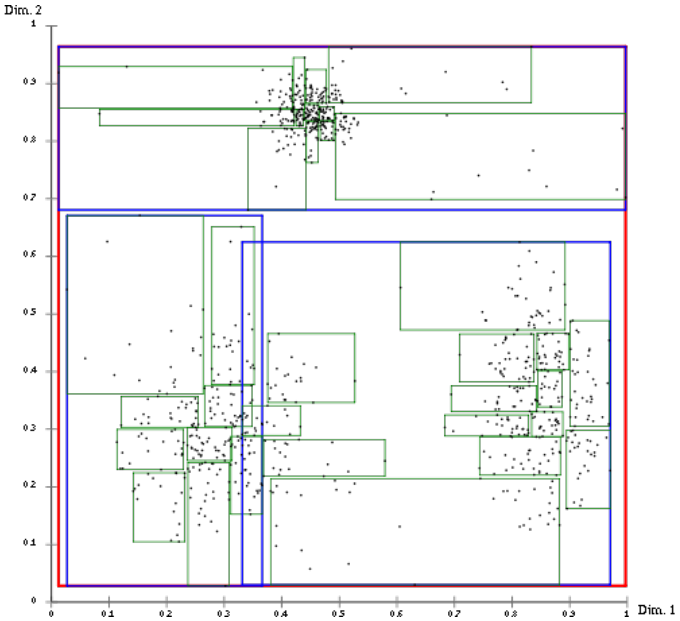
Massive datasets in GRASS 7

Vector processing

Massive datasets in GRASS 7

Vector topology: new spatial index





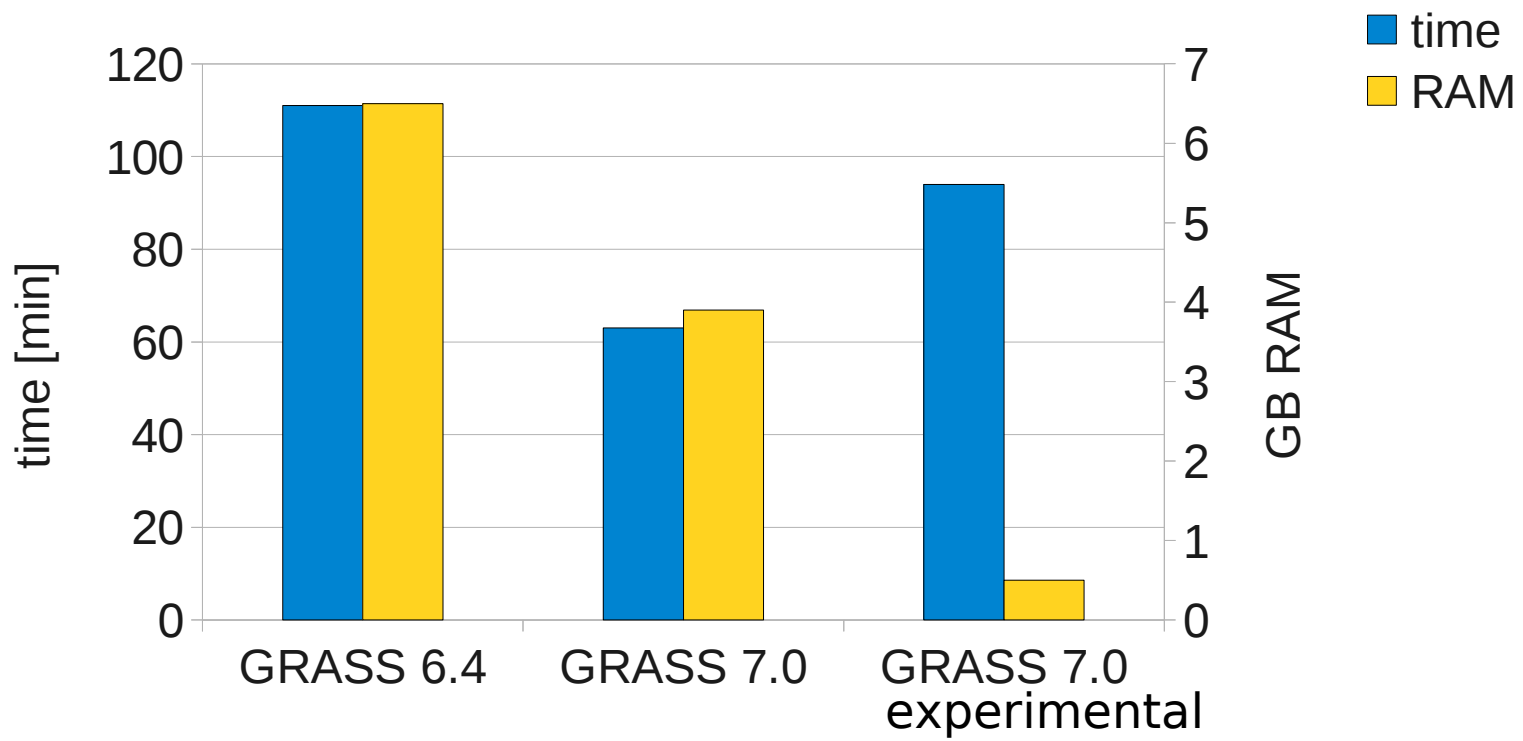
R-Tree

R*-Tree

Massive datasets in GRASS 7

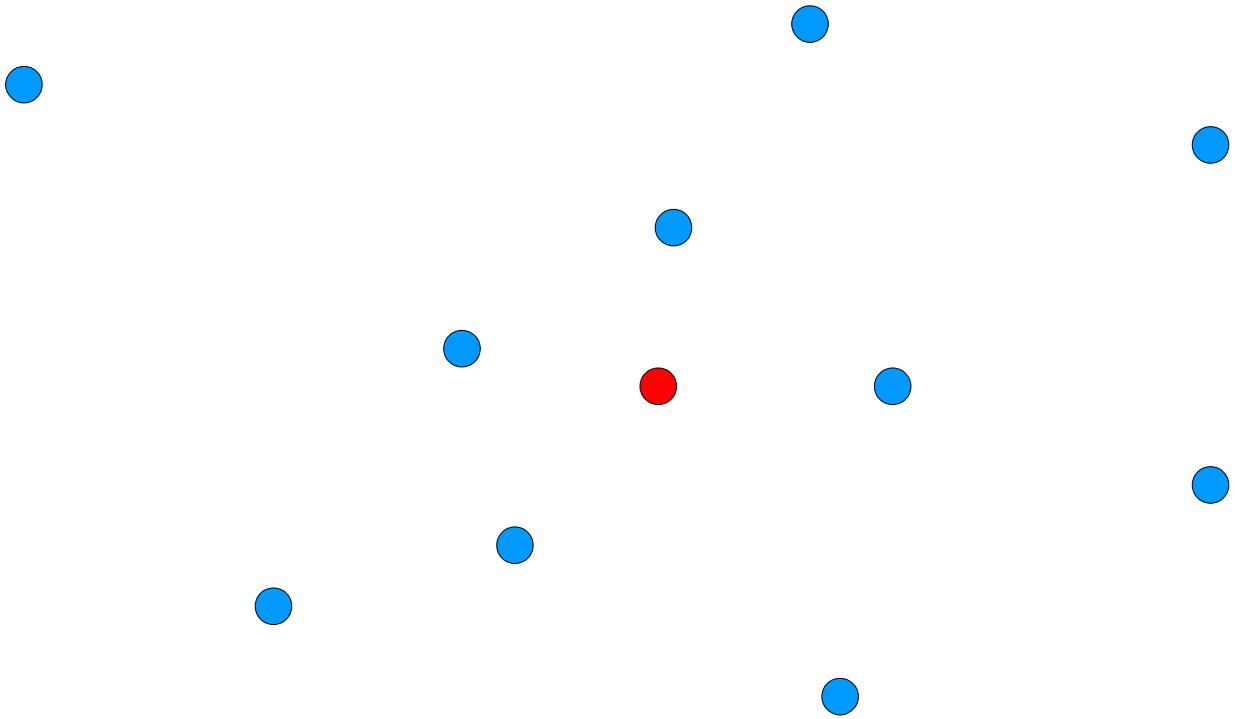
Topological cleaning, e.g. vector polygon import

GADM, all levels, global



Massive datasets in GRASS 7

Vector distance to nearest feature



Massive datasets in GRASS 7

Vector distance to nearest feature

Old:

(n features in from) x (n features in to)

New:

Best case: (n features in from)

Worst case: as old version

Massive datasets in GRASS 7

TODO

experimental -> stable

Further reduce memory requirements (work in progress)

[Bulk of the work: General optimisation of modules]

Massive datasets in GRASS 7

Summary

Raster

Improved external memory for rasters

Selected raster modules speed-up

Vector

Improved vector topology building and cleaning

Thank you