8 May 2020

Analytics Optimized Geoscience Data Store with *STARE-based Packaging*

K-S Kuo[1] and ML Rilee[2]

[1]Bayesics LLC (kuo@bayesics.com)
[2]Rilee Systems Technologies LLC (mike@rilee.net)

What is STARE?

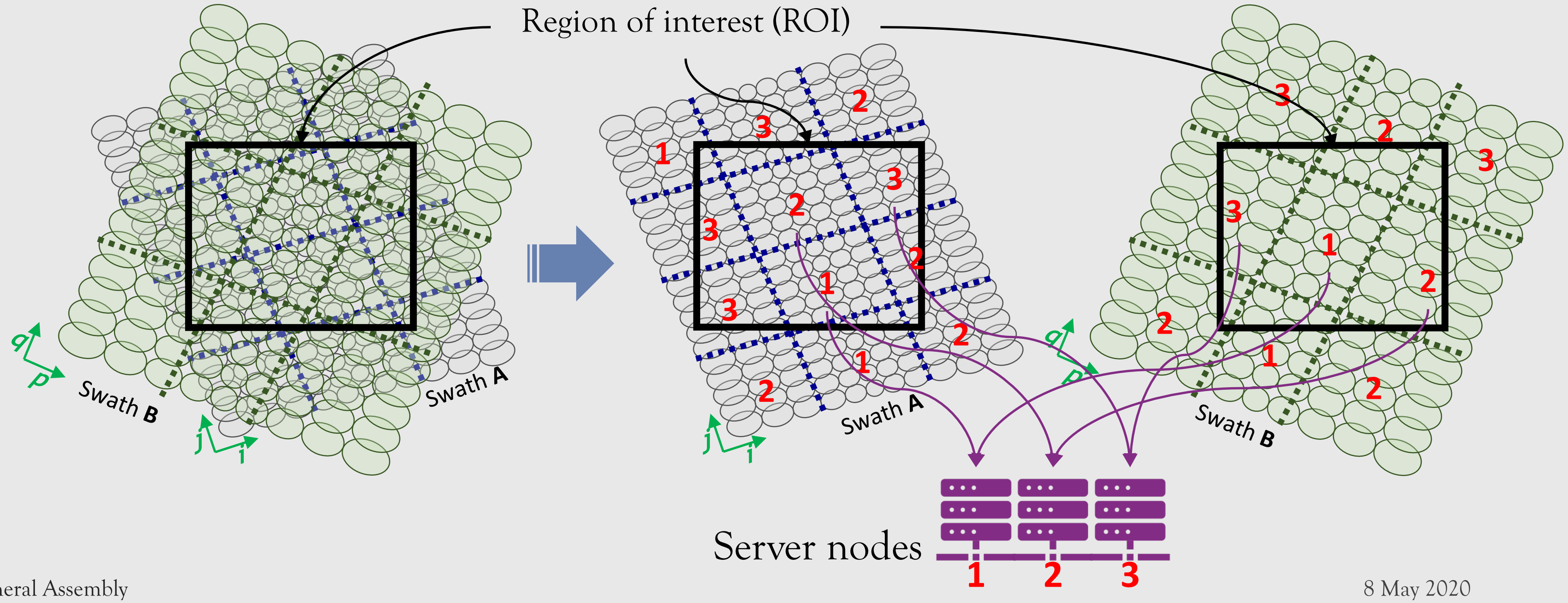
- STARE stands for ***S**patio **T**emporal **A**daptive-**R**esolution **E**ncoding*
- STARE contains two (2) elements:
  - *Spatial* element: HTM - ***H**ierarchical **T**riangular **M**esh*
  - *Temporal* element: HCP – ***H**ierarchical **C**alendrical **P**artition*
- STARE offers a 64-bit integer alternative for indexing geolocation and time
  - Existing geolocation (longitude-latitude) references to data elements use floating-point encoding and indirectly through conventional rectilinear array indices.
- STARE integer indices provide direct geo-spatial and temporal references that are universal across all datasets.
- See *STARE toward unprecedented geo-data interoperability* (Kuo and Rilee 2017).

Why STARE?

- STARE is designed to address data *volume* and *variety* together to meet *velocity* demand.
- *Parallel processing* is the only effective means for scaling volume, but data variety breaks the scalability.
- Data packaging impacts data layout (placement) in memory, which in turn impacts parallel processing efficiency/performance.
- The indirect and ever-changing nature of spatiotemporal references associated with existing data packaging renders *data placement alignment* impossible.
- Misalignments engender unnecessary data movements that reduce scalability.

Example: 2 granules of different orbits and resolutions

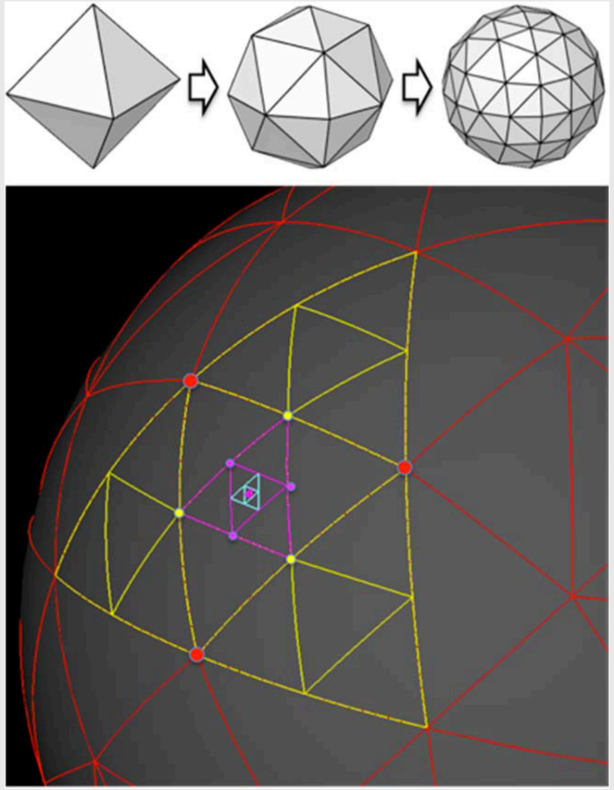
- It is <u>impossible</u> to align data placements *spatiotemporally* onto server nodes using conventional array indices!
- But, most geoscience data analyses require processing diverse geo-data *for the same place and time*.
- When data are misaligned in memory, communications among nodes ensue and break scalability.

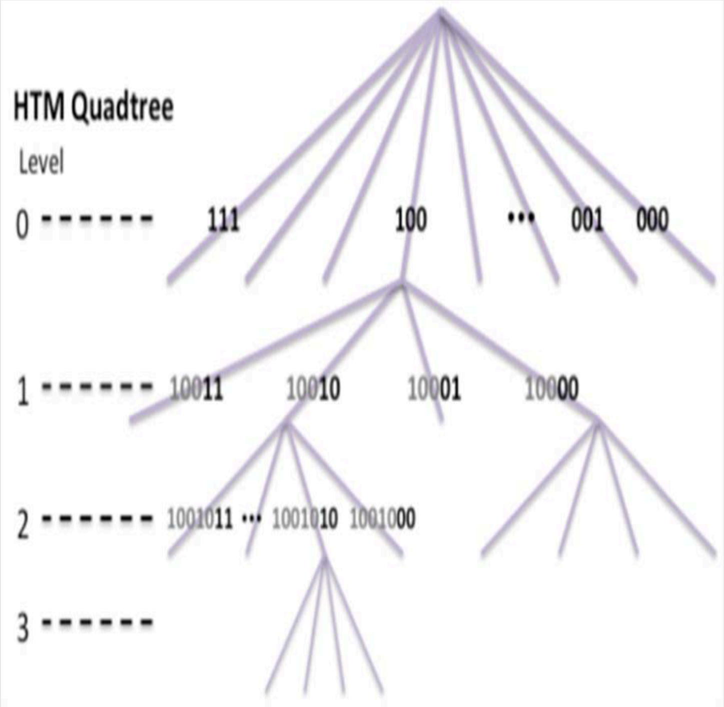


STARE Hierarchical Triangular Mesh

HTM indexes the surface (actually, solid angle) of a sphere using a hierarchy of spherical triangles.

1. Start with an inscribing *octahedron* of a sphere.
2. Bisect each edge.
3. Project the bisecting points from sphere center to the sphere surface to form 4 smaller spherical triangles (quadfurcation).
4. Repeat from 2.

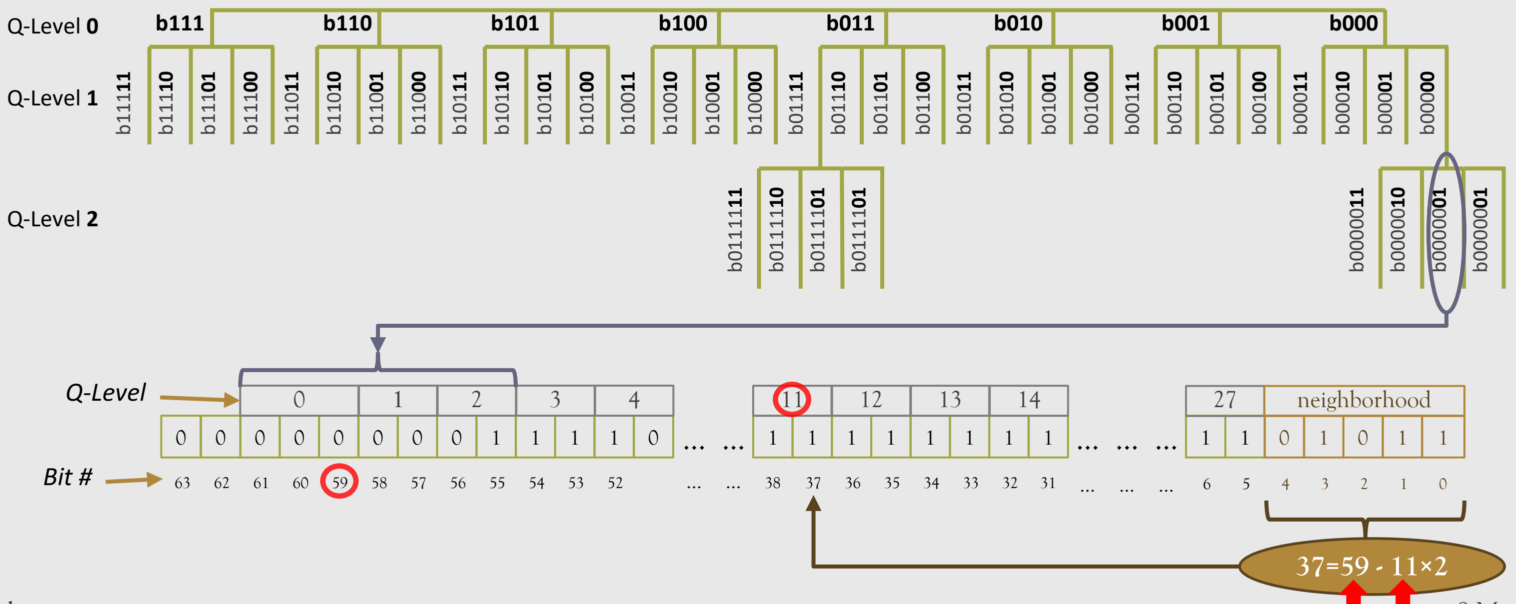




STARE HTM Index Bit-format

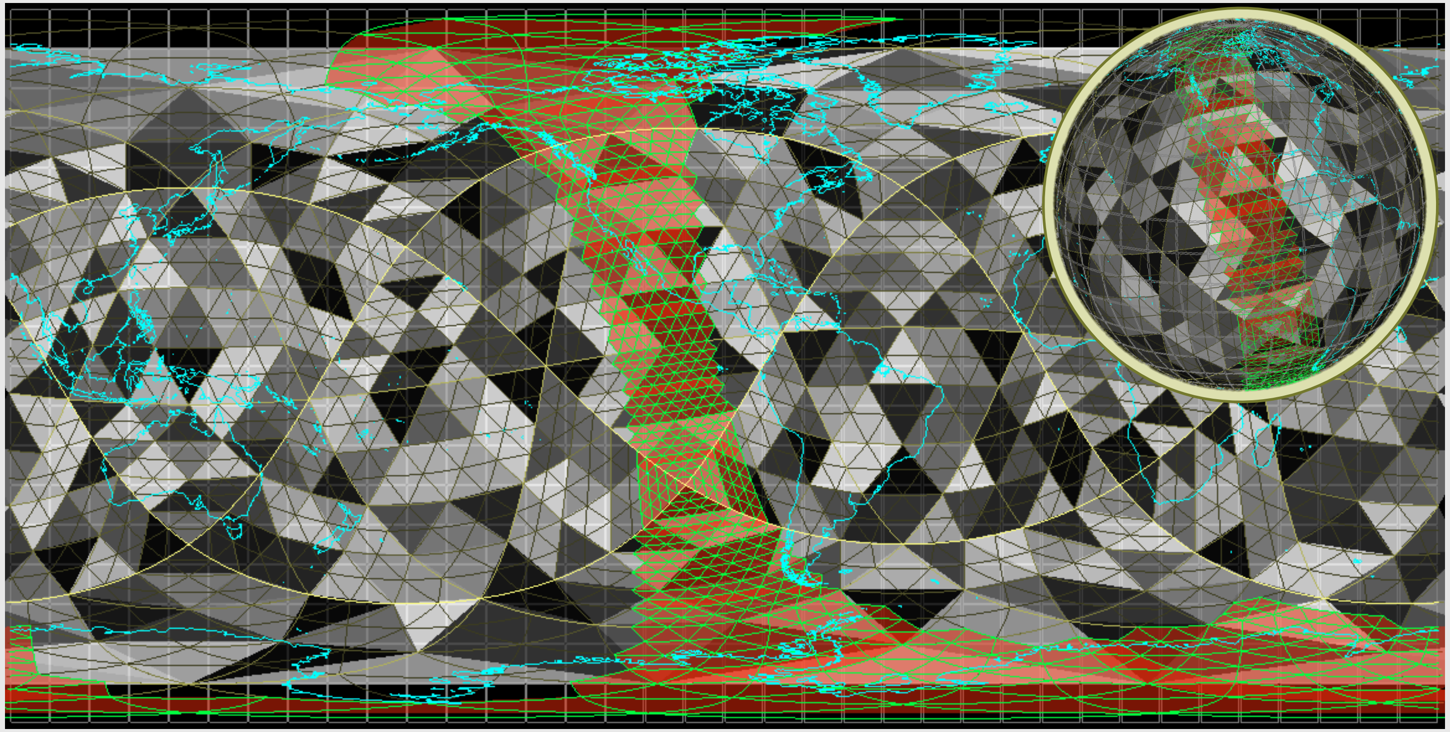
STARE HTM index includes both geolocation and resolution information in a 64-bit integer:

- Most of the more significant bits (59) are used to indicate geolocation to ~8 cm precision.
- The least significant 5 bits are are used to indicate the quadfurcation level (QL) with mesh size close to the data resolution.



Partitions of Earth Surface at a Quadfurcation Level

- If all data are indexed with STARE indices, all can be partitioned the same way.
- When all data are partitioned the same way, it guarantees spatiotemporal co-alignment of data on storage or in memory.
- When data are aligned, unnecessary data movements are minimized.



Furthermore…

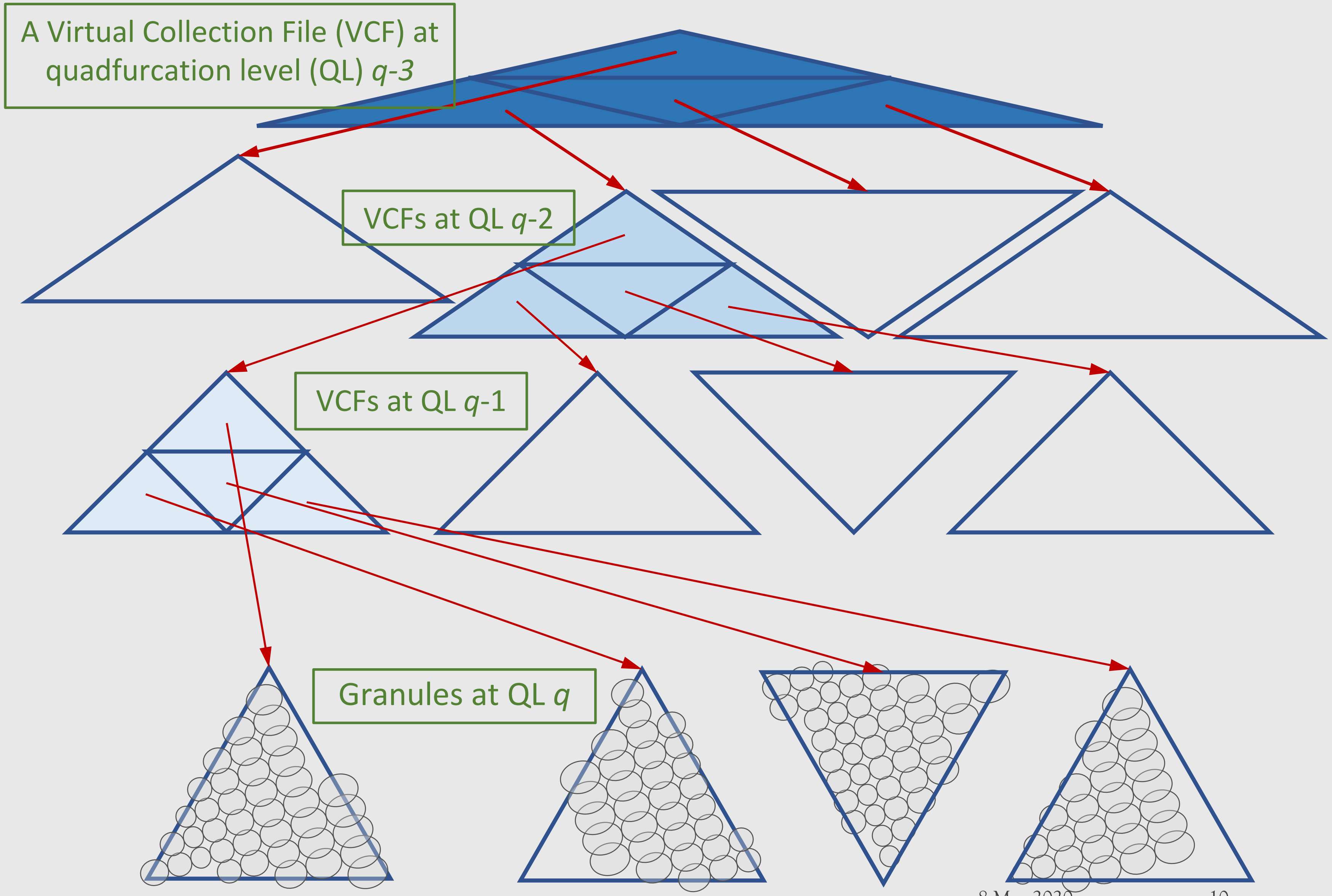
STARE indexes diverse geo-data in their native forms, without requiring *re-gridding* to regular longitude-latitude grid or raster format.

- It reduces the need for intermediate data and thus total data volume.
- Scientists prefer working with "*unaltered*" data.
- Re-gridding is left as a step in a highly scalable process, in which the scientist can choose a re-gridding method per analysis requirement.
- Traceability to source/original data is much improved.

STARE-based Packaging

- Existing packaging predominantly stores geo-data as conventional rectilinear arrays, making spatiotemporal data placement alignment exceedingly difficult.
- Leveraging STARE, geo-data can be packaged hierarchically, enabling straightforward spatiotemporal data placement alignment.

- Determine a quadfurcation level (QL), *q*, to package "physical" granules.
- Virtual granules for QL's above *q* can be formed hierarchically, effectively providing a range of partition sizes.
- Analysis techniques can choose an appropriate QL, depending on resource availability, to read and align granules from diverse geo-data.



Conclusions

- The universality of STARE integer indices effectively harmonizes geo-data variety amid volume.
- The hierarchical nature of STARE indices uniquely supports flexible geo-data partition and spatiotemporal data placement alignment.
- STARE-based packaging makes it easier to take advantage of STARE potentials.
- STARE-based packaging is expected to afford unparalleled scalability with geo-data interoperability.

Acknowledgments & References

- STARE is available at https://github.com/SpatioTemporal
  - *The STARE C++ base library*
  - *PySTARE* is an API for Python with GeoPandas functions
  - *STARE-Cookbooks* contains a growing number of examples
- STARE development is supported by NASA Advancing Collaborative Connections for Earth System Science (ACCESS-17) Award No. 80NSSC18M0118.