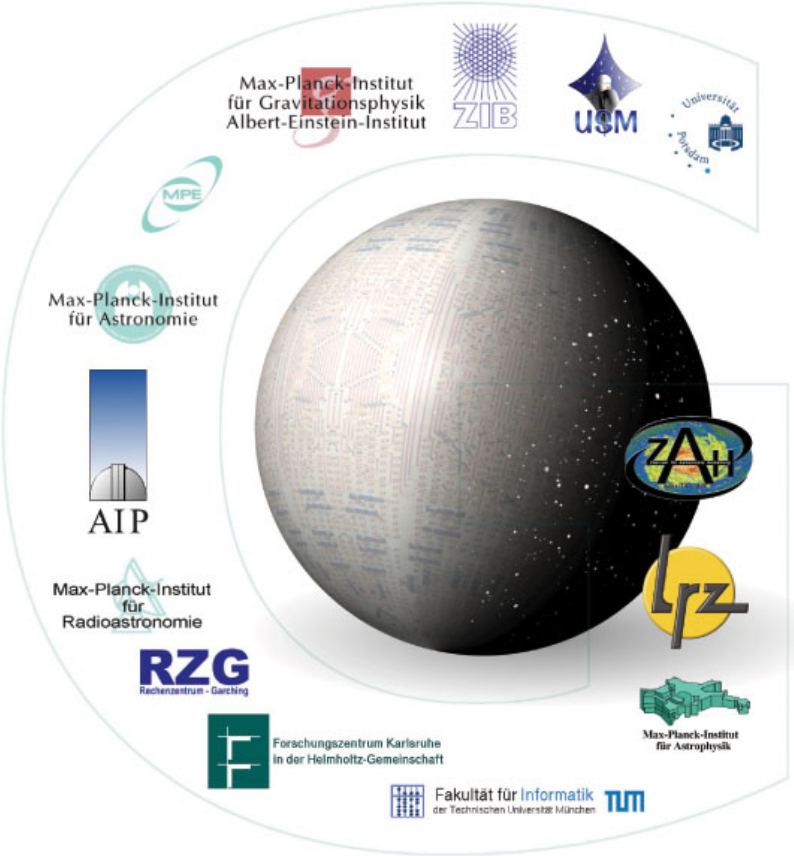
Community Training: Partitioning Schemes in Good Shape for Federated Data Grids

Tobias Scholl, Richard Kuntschke, Angelika Reiser, Alfons Kemper

3rd IEEE International Conference on e-Science and Grid Computing
Bangalore, India
December 10th – 13th 2007

The AstroGrid-D Project

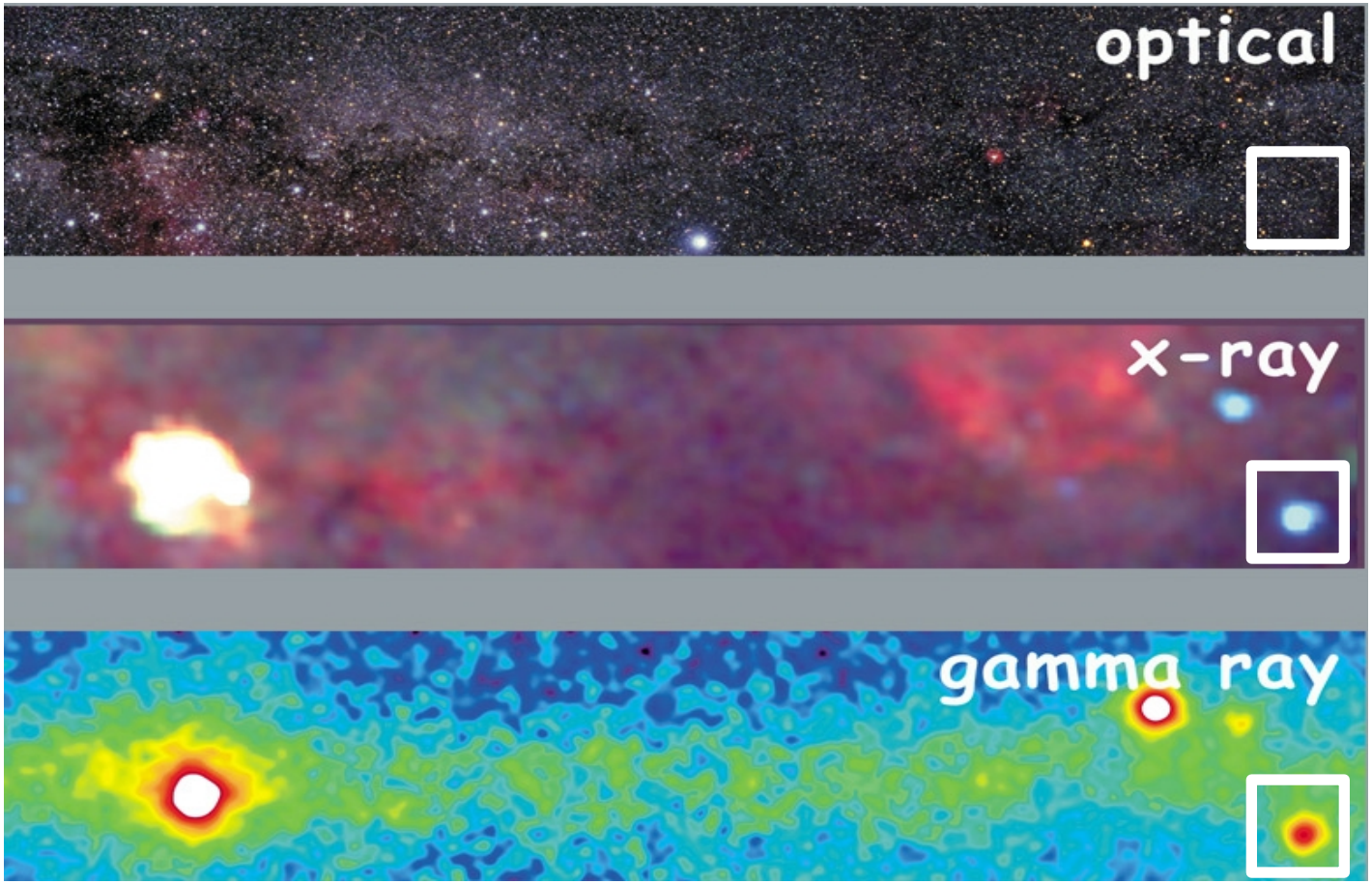
- German Astronomy Community Grid http://www.gac-grid.org/
- funded by the German Ministry of Education and Research
- part of the D-Grid



Federal Ministry of Education and Research

D-GRID

The Multiwavelength Milky Way



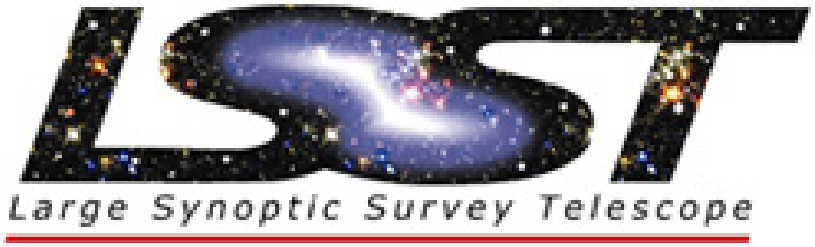
http://adc.gsfc.nasa.gov/mw/

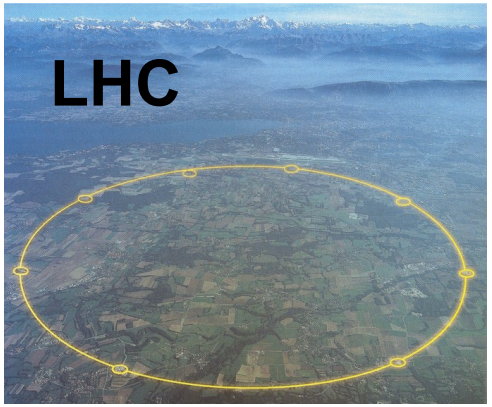
Data-intensive e-Science Applications

- Many e-science application areas
  - astrophysics
  - geosciences
  - climatology
- Combination of various, globally distributed data sources
- Increasing popularity (within community and public domain): more users
- Scalability issues with current approaches

Up-coming Data-intensive Applications

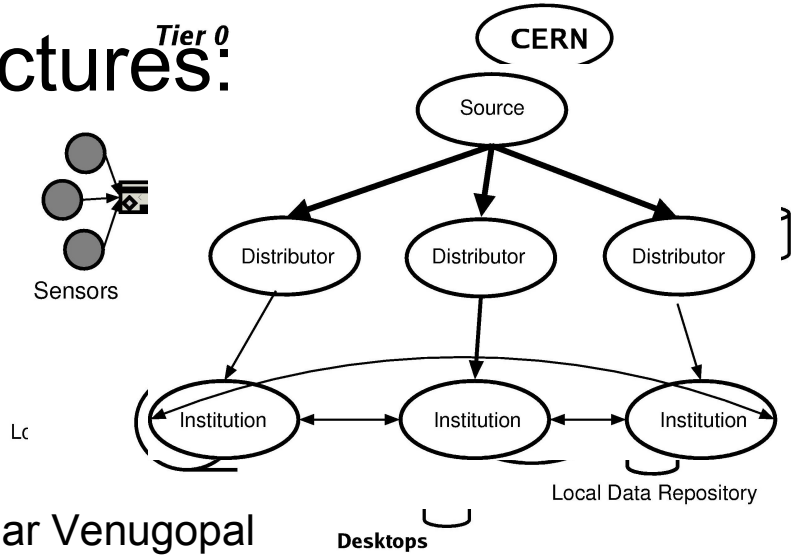
- Data rates
  - Terabytes a day/night
  - Petabytes a year
- LOFAR
- LSST
- Pan-STARRS
- LHC





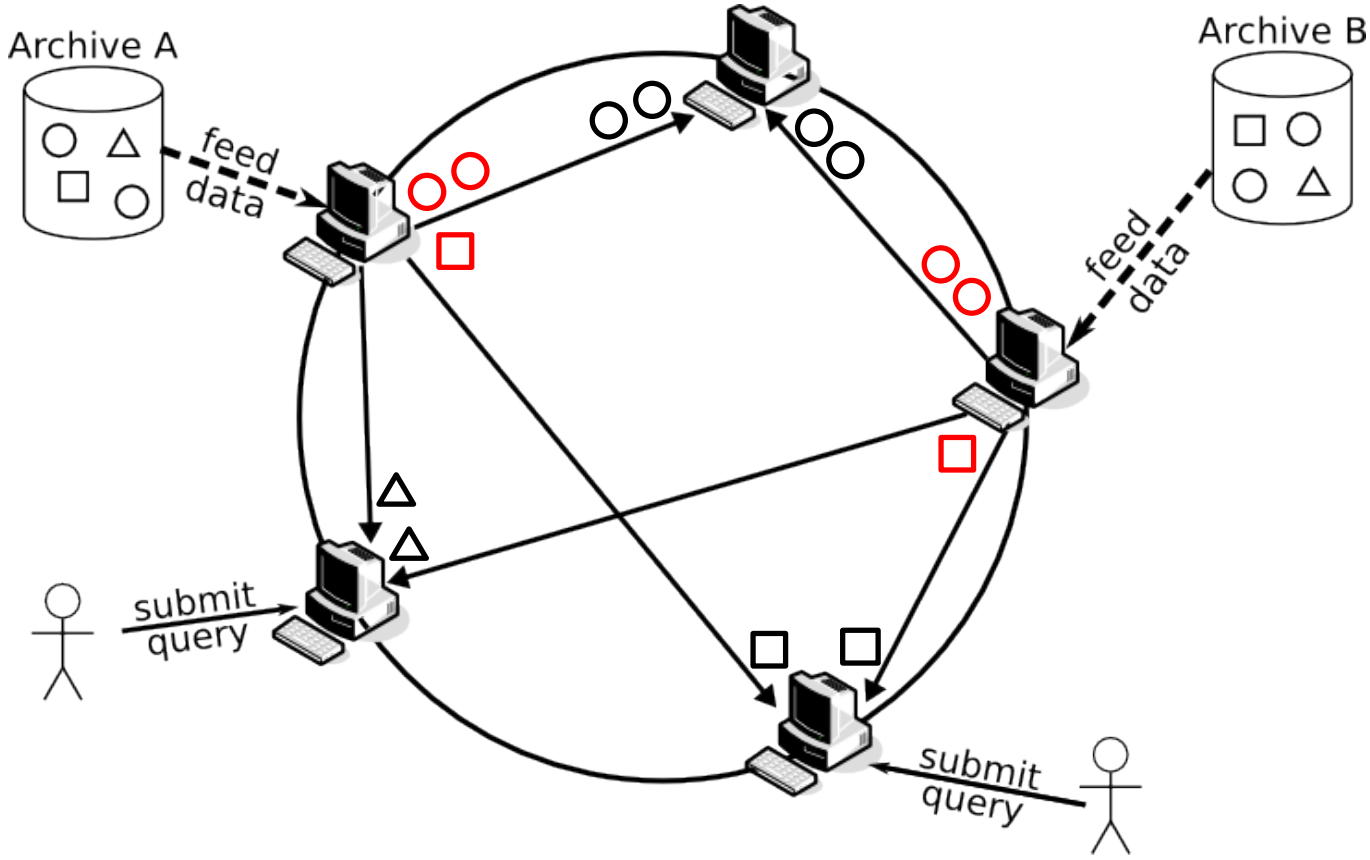
Current Sharing in Data Grids

- Data autonomy
- Policies allow partners to access data
- Each institution ensures
  - Availability (replication)
  - Scalability
- Various organizational structures:
  - Centralized
  - Hierarchical
  - Federated
  - Hybrid



Images from: “Data-Intensive Grid Computing” by Srikumar Venugopal

Community-Driven Data Distribution

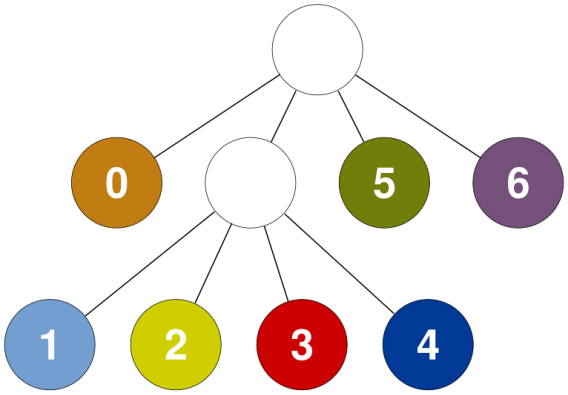


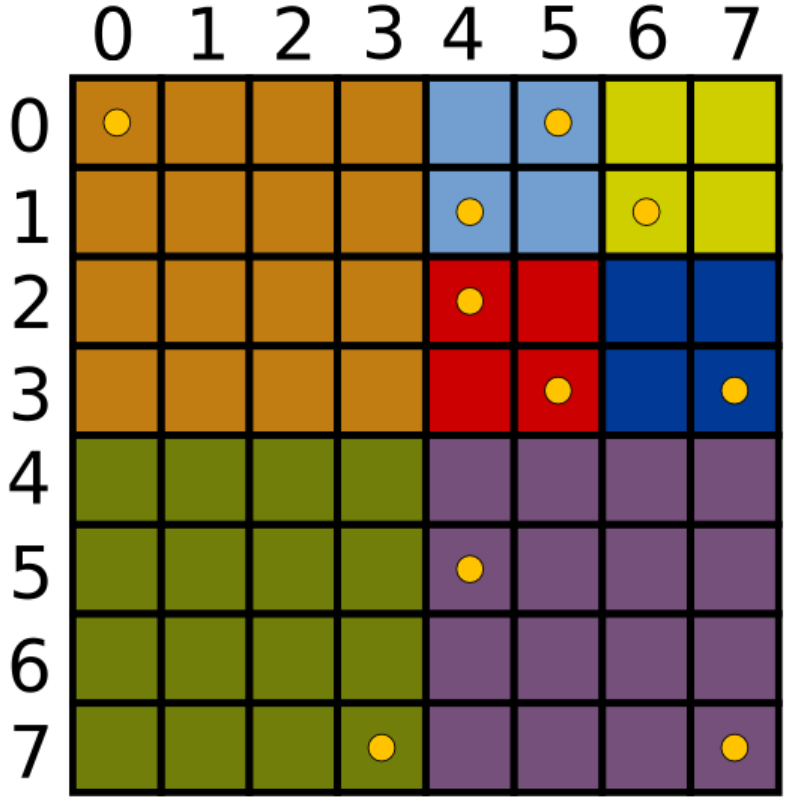
The Training Phase

- Extract data from the archives
- Compute partitioning schemes
- Compare different partitioning schemes
  - Standard Quadtrees
  - Median-based Quadtrees
  - Zones

Quadtrees

- Well-known index structure
- Recursive decomposition
- Adaptive to data resolution





Quadtree-based Schemes: Splitting Variants

Center splitting

- Always bisects each dimension
- congruent subareas
- Splitting points stored or computed

Median heuristics

- Compute median for each dimension independently
- O(n) median algorithm
- Splitting points stored

The Zones Index (J. Gray, M. Nieto-Santisteban, A. Szalay)

- Index structure for databases
- Specific spatial clustering in zones
- Optimized for
  - points-in-region queries
  - self-match, cross-match queries
- Equi-distant partitioning
  - Declination coordinate
  - Zone(ra, dec) = floor((dec + 90.0) / h)
- Implemented directly in SQL

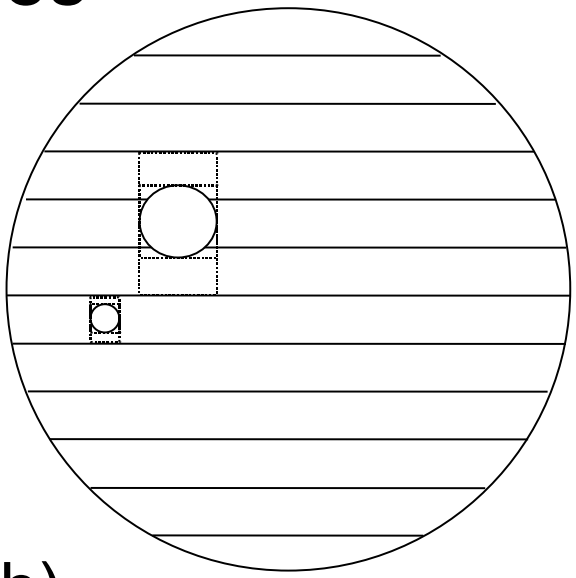
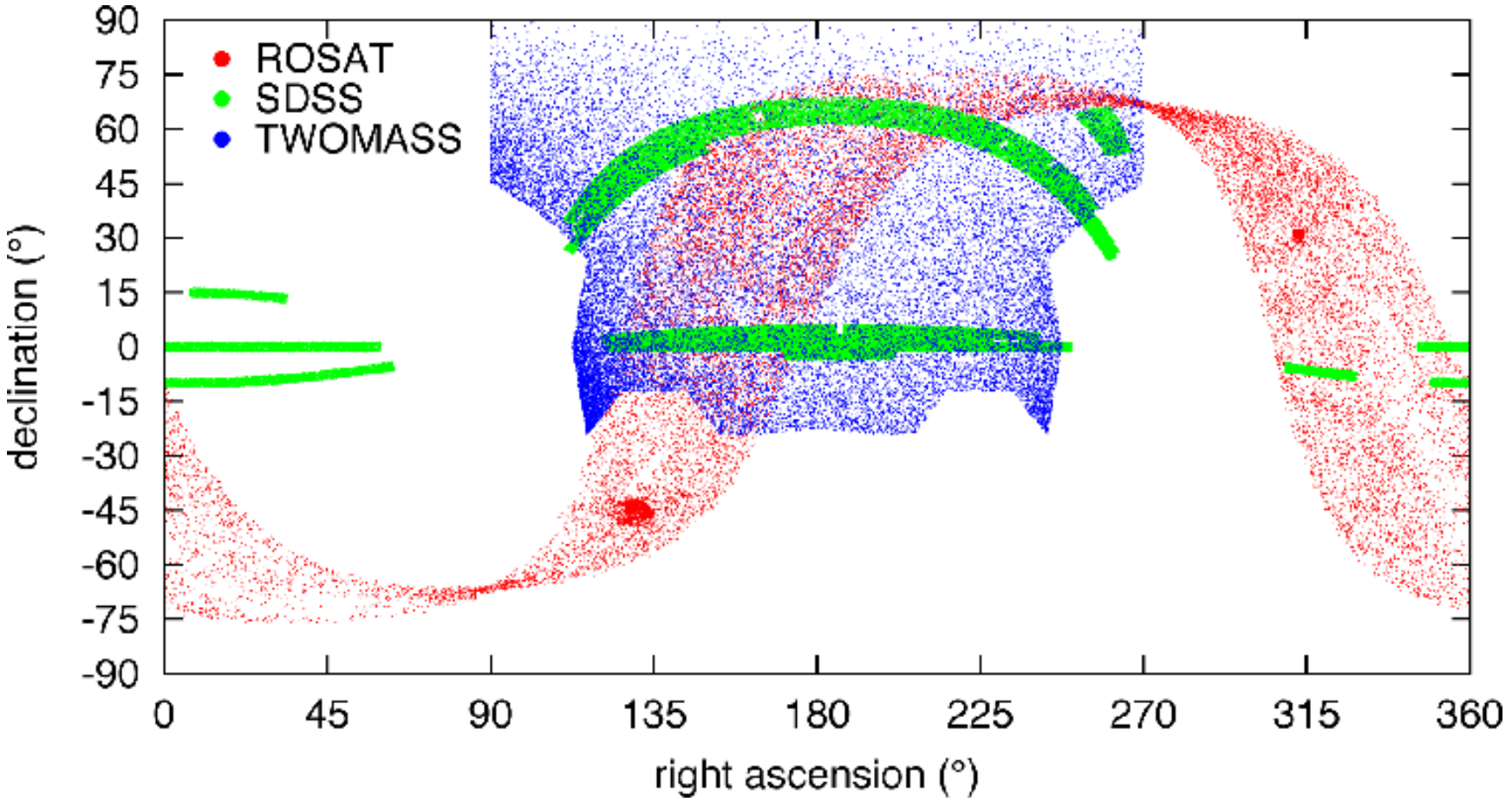


Image by J. Gray et al.

Evaluation Setup

- 2 data sets: skewed and uniform
- Size of data sample: 0.01%, 0.1%, 1%, 10%
- Number of partitions:
  4, 16, 64, 256, 1024, 4096, 8192, 16384, 32768, 65536, 131072 ($2^4 – 2^{17}$)

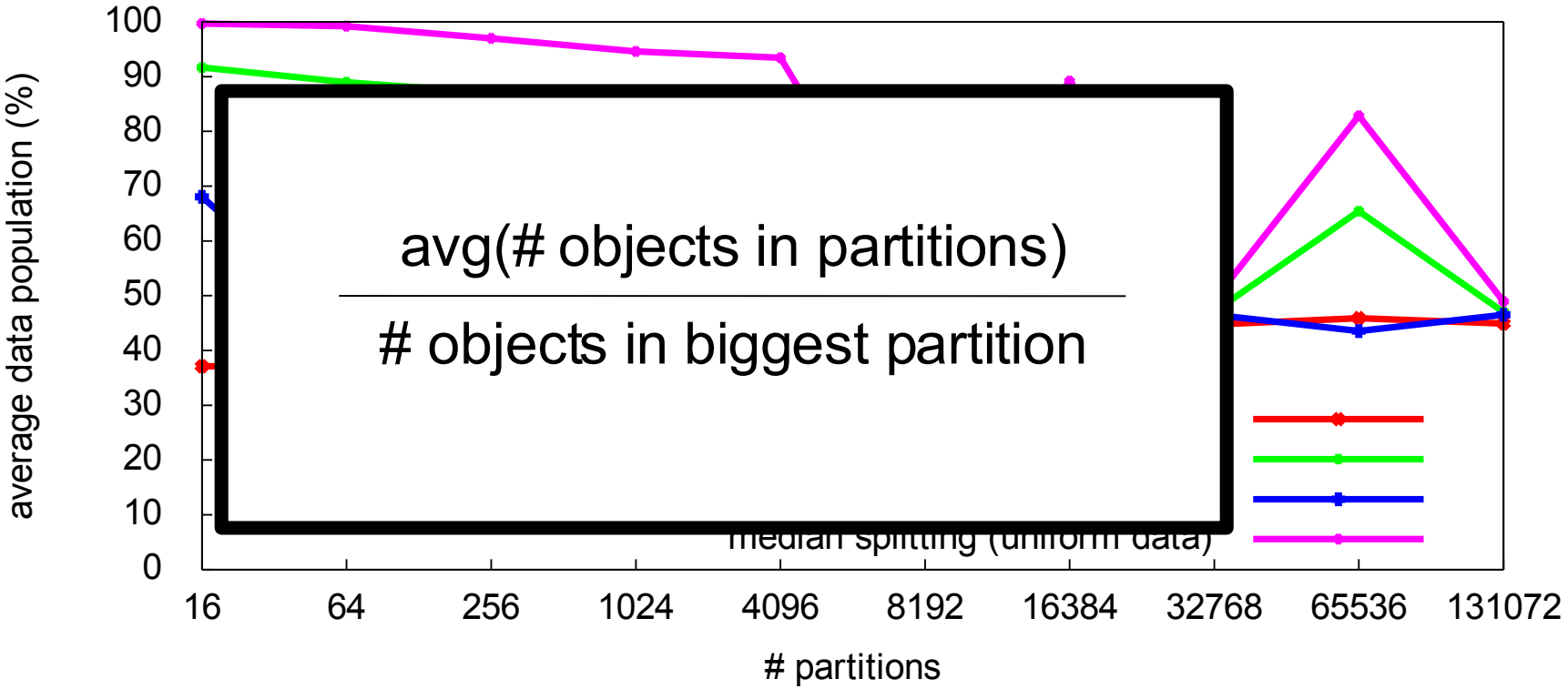
Skewed Training Data (D_{skew})



Comparing Partitioning Schemes

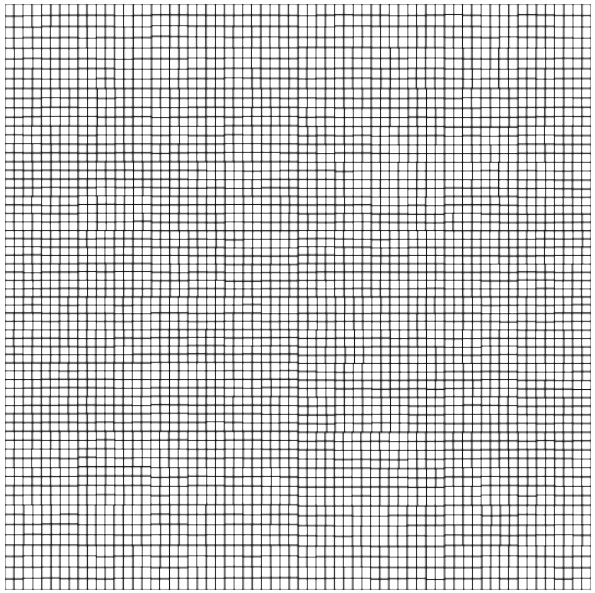
- Duration
- Average data population
- Variance in partition population
- Empty partitions
- Size of the training set
- Baseline comparison

Average Data Population

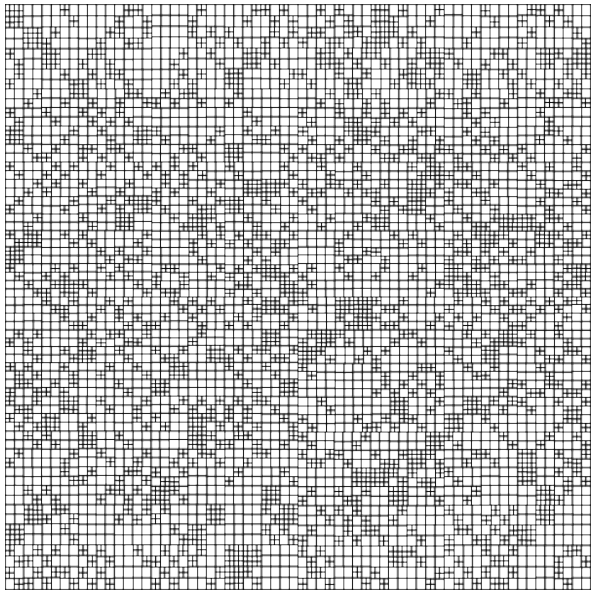


10% training sample, D_{skew}

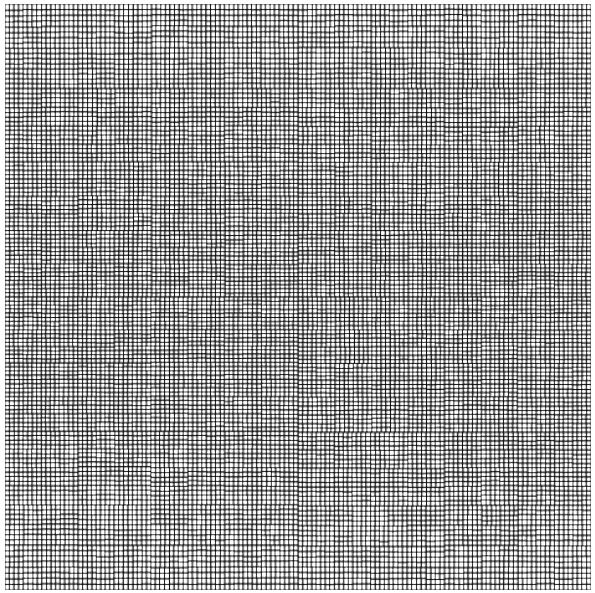
Evolution of the Partitioning Scheme



4096 partitions

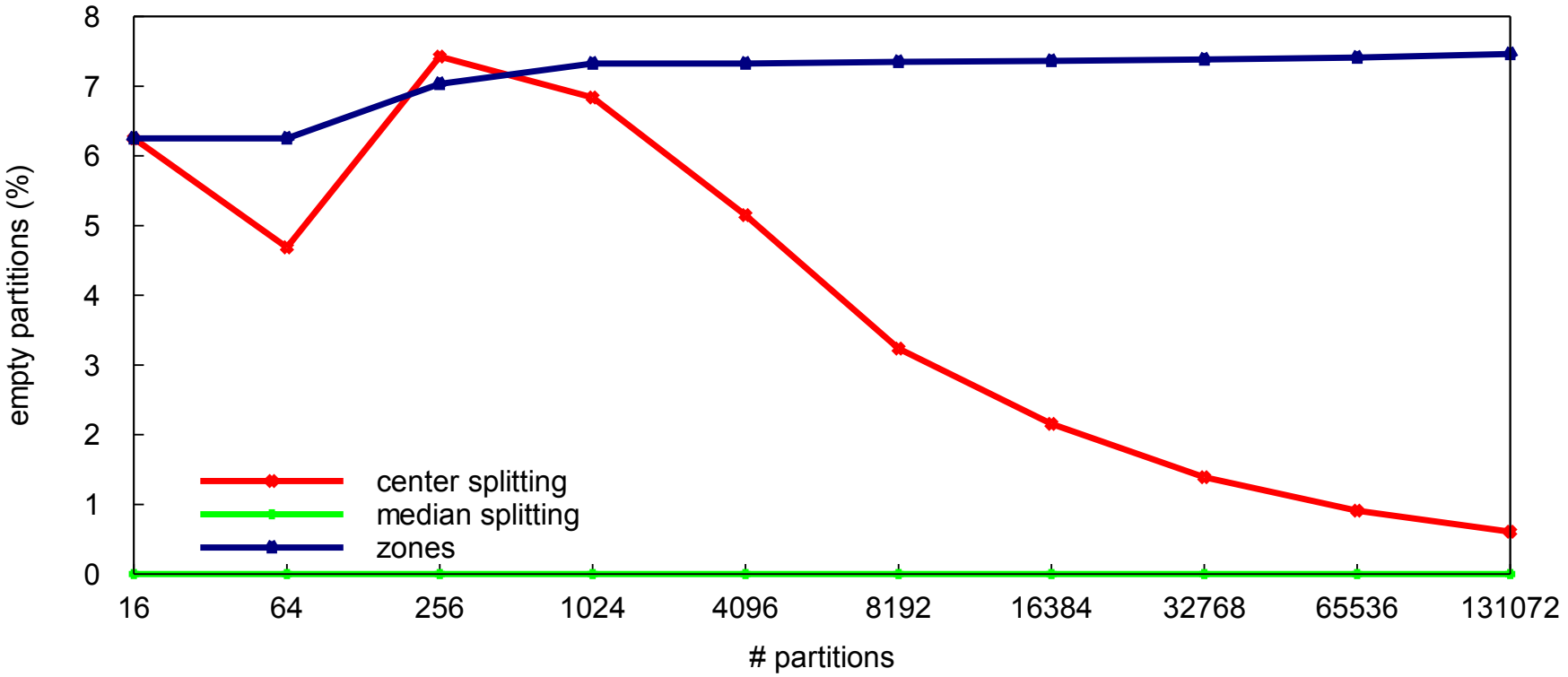


8192 partitions



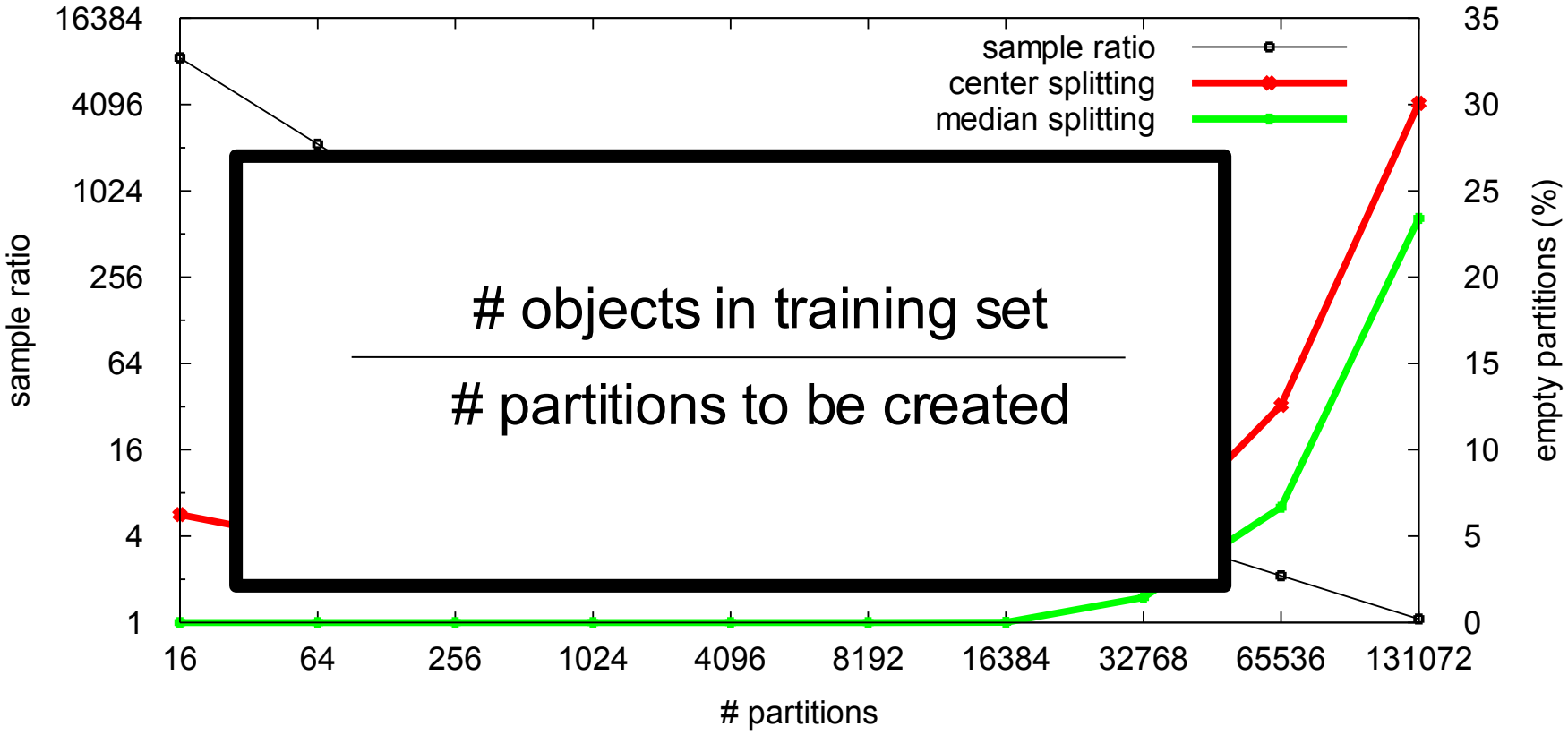
16384 partitions

Empty Leaves



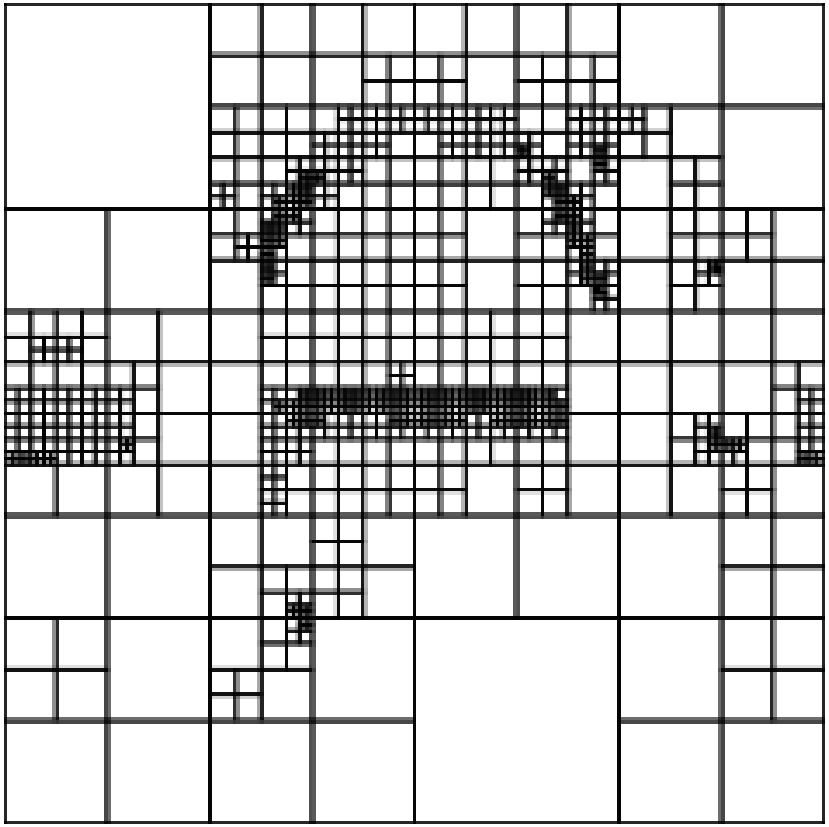
10% training sample, D_{skew}

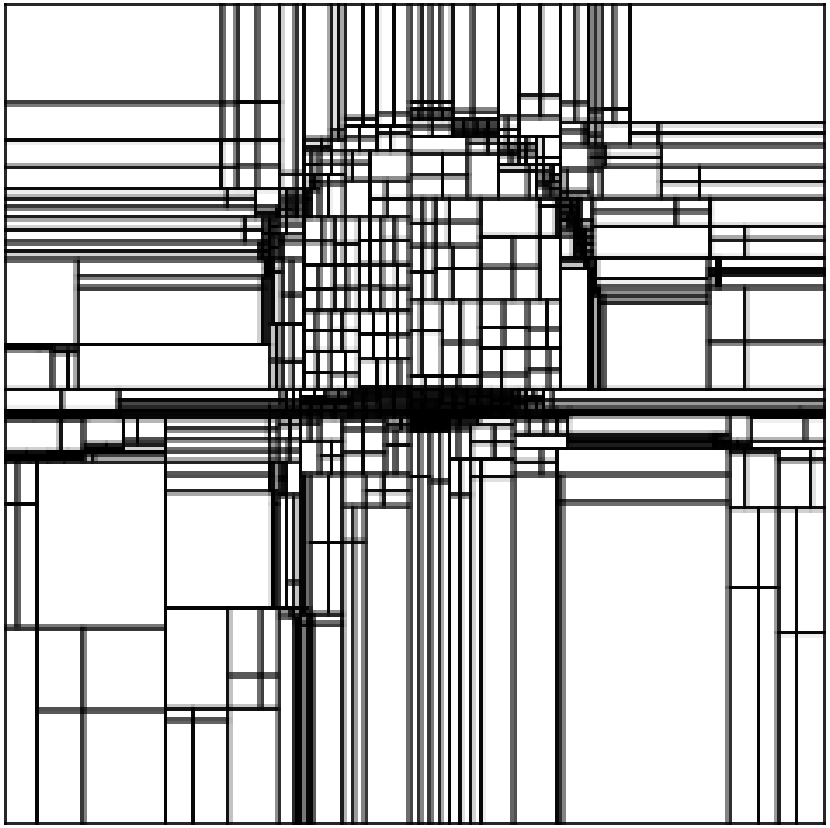
Size of the Training Set



0.1% training sample, D_{skew}

Standard Quadtree vs. Median Heuristics



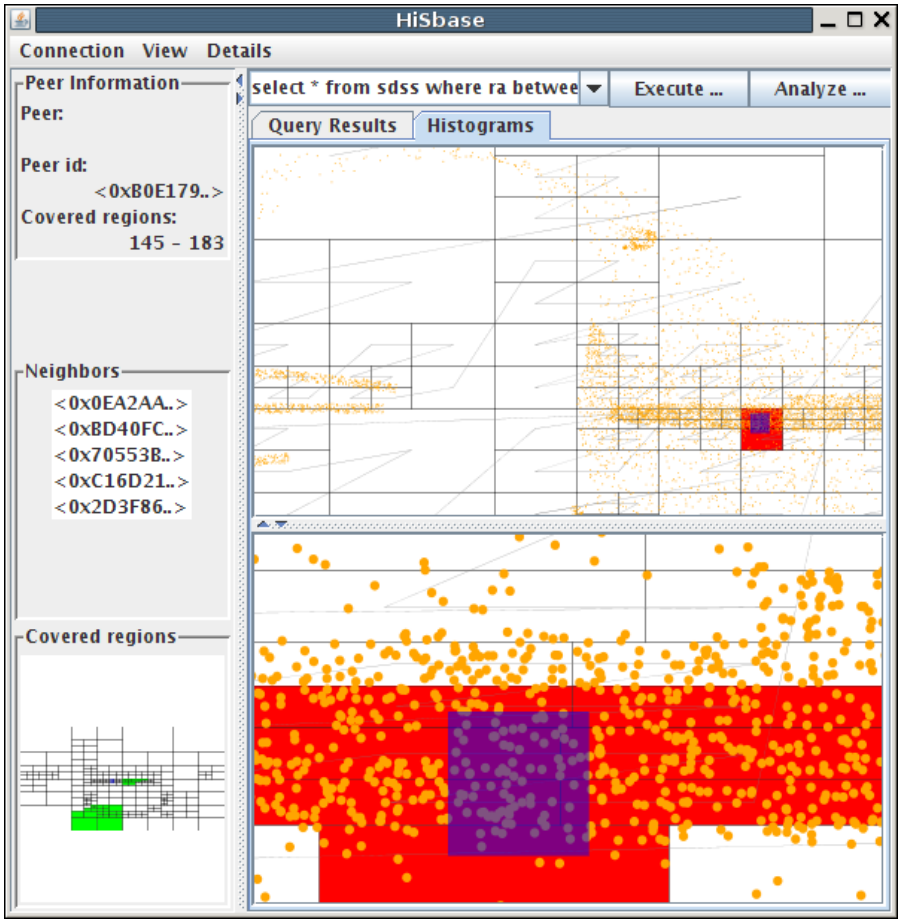


Evaluation – Summary

- Quadtrees: good adaption to data distribution
- Quadtrees: Trade-off between data load balancing and uniformly shaped regions
- Median-based heuristics: best data load balancing even for skewed data sets
- Zone Index: good for uniform data sets
- Training set needs to be sufficiently large in order not to artificially create empty partitions

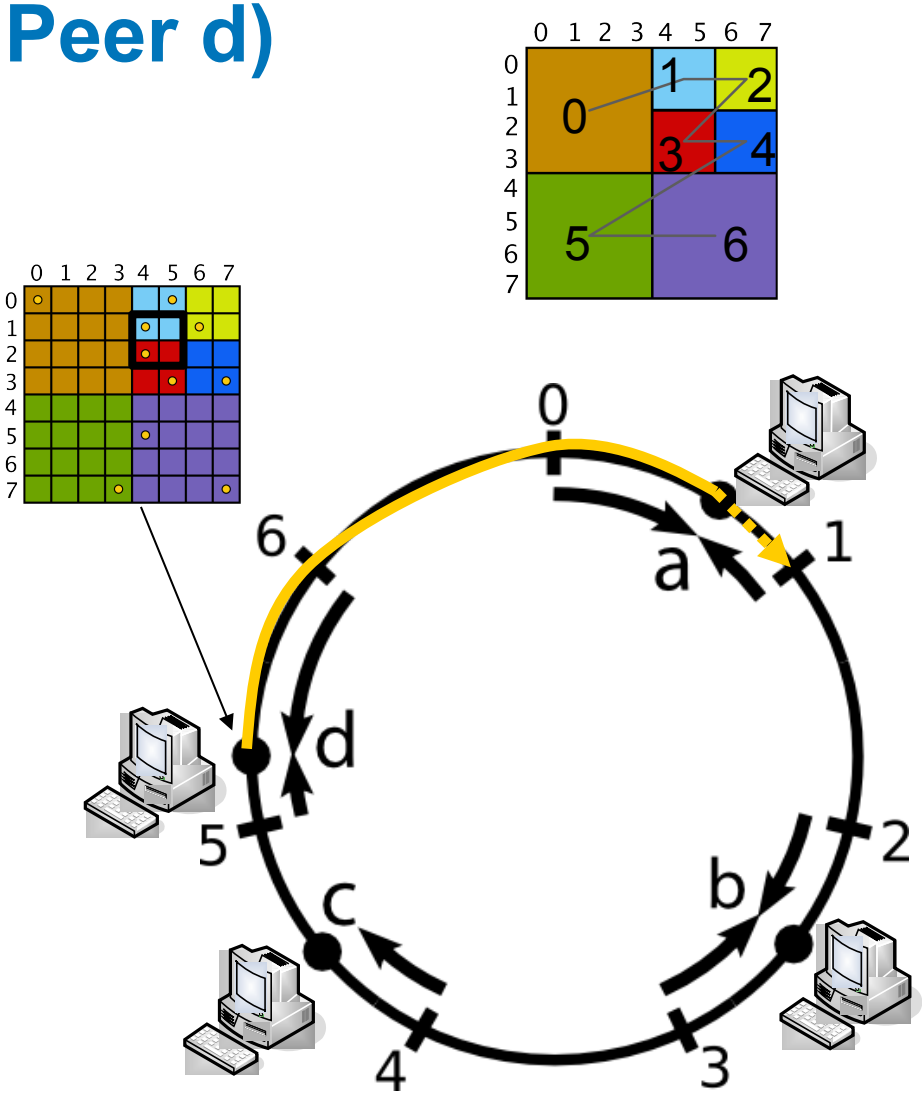
HiSbase

- Peer-to-Peer layer assigns data partitions to peers
- Higher flexibility
- New peers are integrated seamlessly



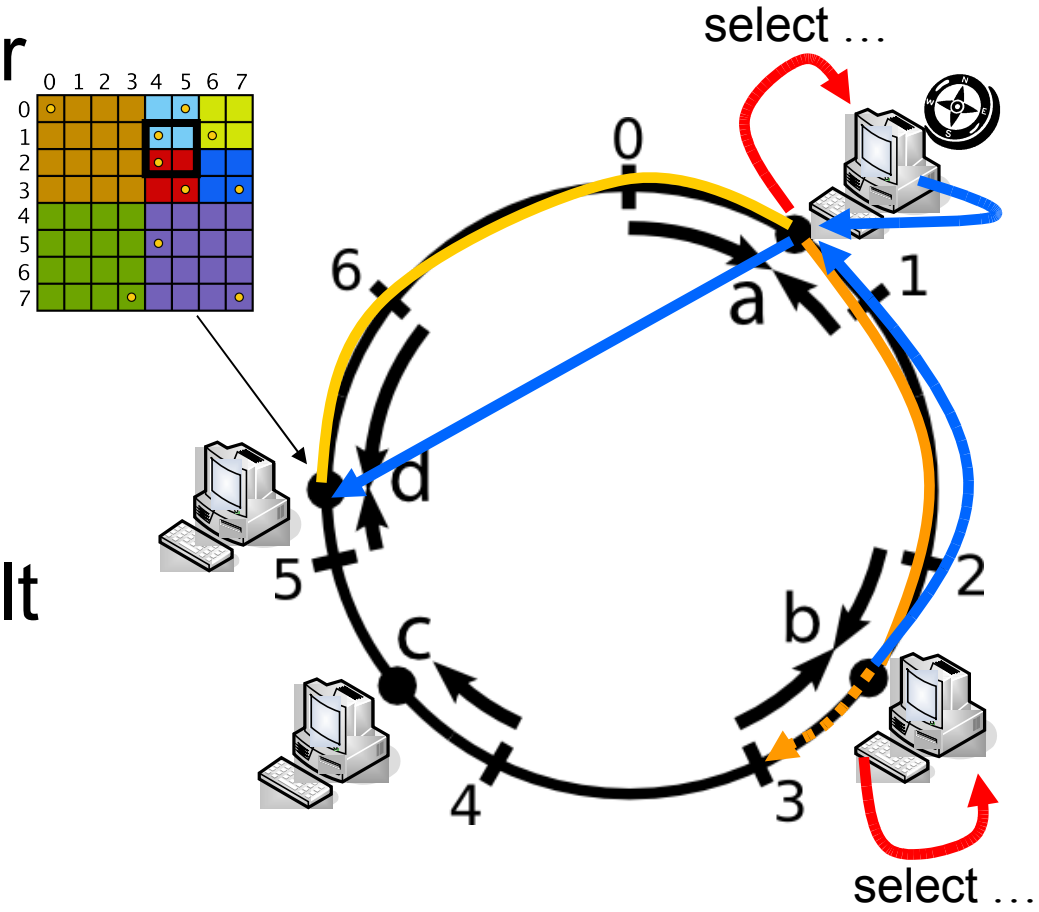
Query Submission (at Peer d)

- Determine relevant regions: [1,3]
- Select coordinator: Region 1
- Send CoordinateQuery-message to id1: {[1,3], SQL}
- Message gets routed to Peer a.



Query Coordination (at Peer a)

- Peer a is coordinator
- Contact relevant regions
- Collect intermediate results
- Send complete result to client



Prototype Implementation

- Java-based prototype
- FreePastry library (Pastry implementation)
  - Rice University
  - MPI-SWS
- Presentations at
  - BTW 2007, Aachen, Germany
  - VLDB 2007, Vienna, Austria
- Deployed in various settings
  - LAN
  - WAN (AstroGrid-D, PlanetLab)

Summary

- Training phase
- Community-driven Data Grids
  - Domain-specific partitioning scheme
  - Partitioning scheme supports
    - Data skew
    - Region-based queries
- Framework for comparing partitioning schemes
- Various measures with regard to data load balancing

Ongoing Work

- Database-driven comparison
  - 0.1% of a Petabyte still is 1 Terabyte
  - Feasibility of median-based techniques
- Workload-aware data partitioning
- Heterogeneous data nodes

Get in Touch

- Database systems group, TU München
  - Web site: http://www-db.in.tum.de
  - E-mail: scholl@in.tum.de
- HiSbase
  - http://www-db.in.tum.de/research/projects/hisbase/
- Data stream management
  - “Grid-based Data Stream Processing in e-Science” (e-Science '06)
  - http://www.gac-grid.de/project-products/Software/DataStreamManagement.html

AstroGrid-D Research Demo

- Finding Galaxy Clusters using Grid Computing Technology
- Room: Banquet
- Wednesday 3:30 pm to 6:00 pm