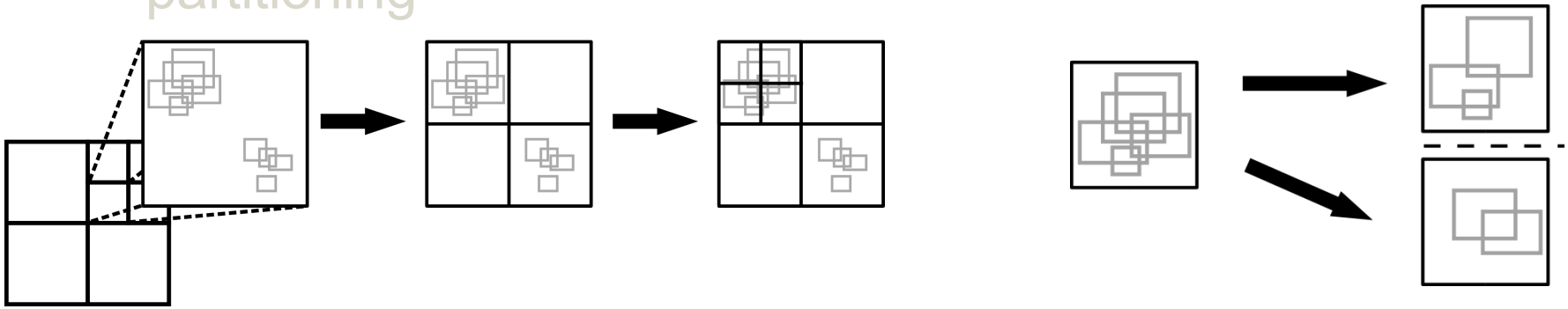
Workload-Aware Data Partitioning in Community-Driven Data Grids

Tobias Scholl, Bernhard Bauer, Jessica Müller, Benjamin Gufler, Angelika Reiser, and Alfons Kemper

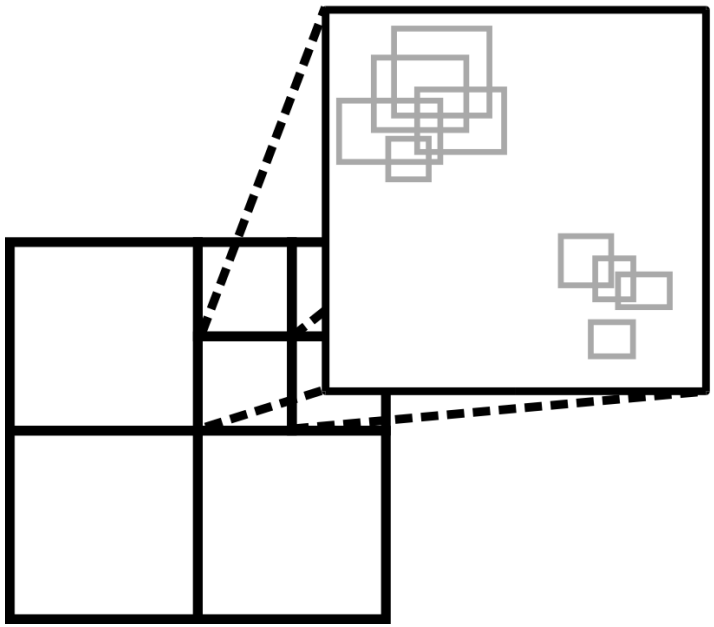
Department of Computer Science, Technische Universität München
Germany

- Many challenges and opportunities in e-science for database research
  - High-throughput data management
  - Correlation of distributed data sources
- Community-driven data grids
  - Dealing with data skew and query hot spots
  - Workload-awareness by employing cost model during partitioning

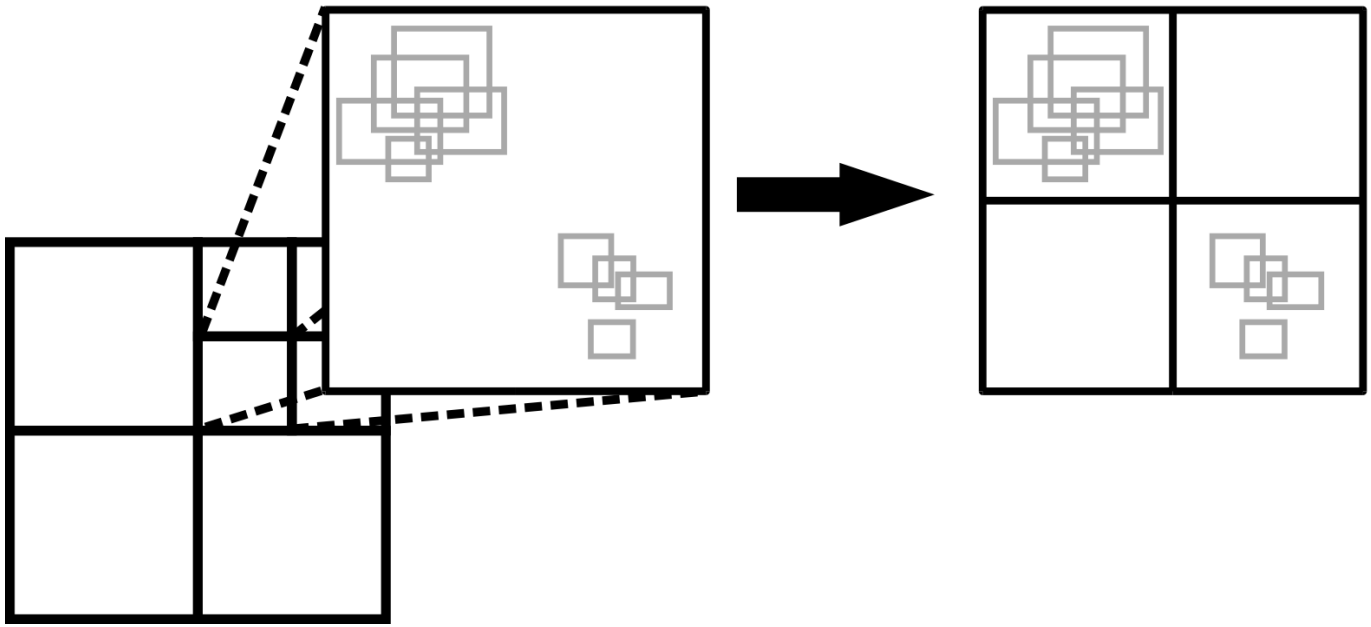
Should I Split or Replicate?



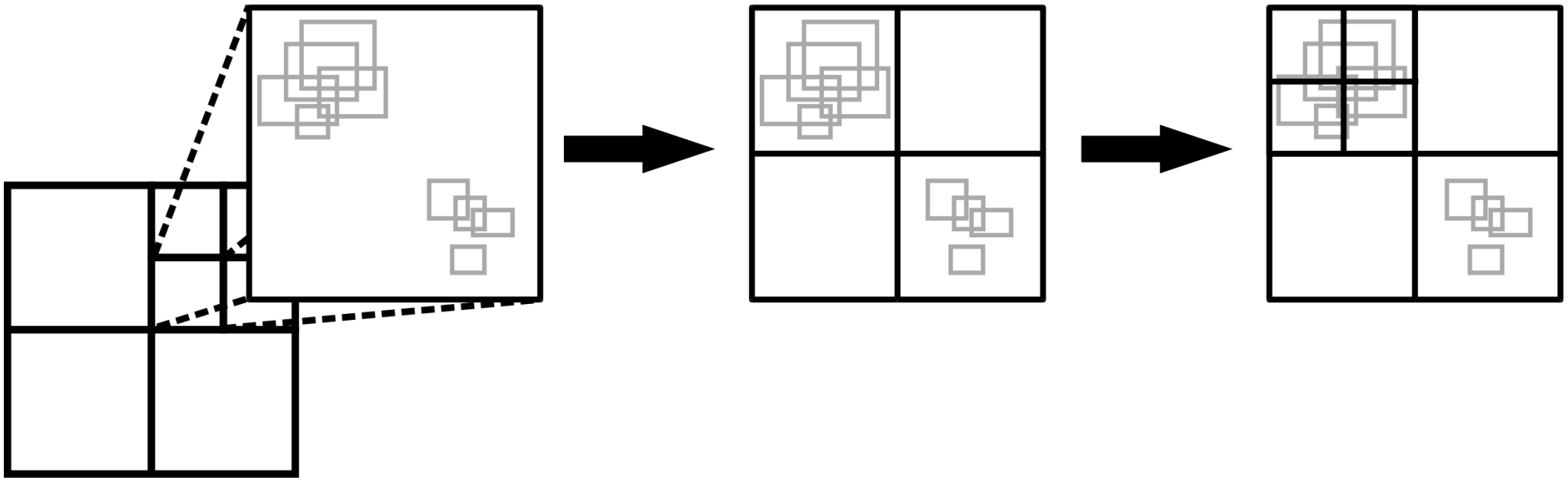
Query Load Balancing via Partitioning



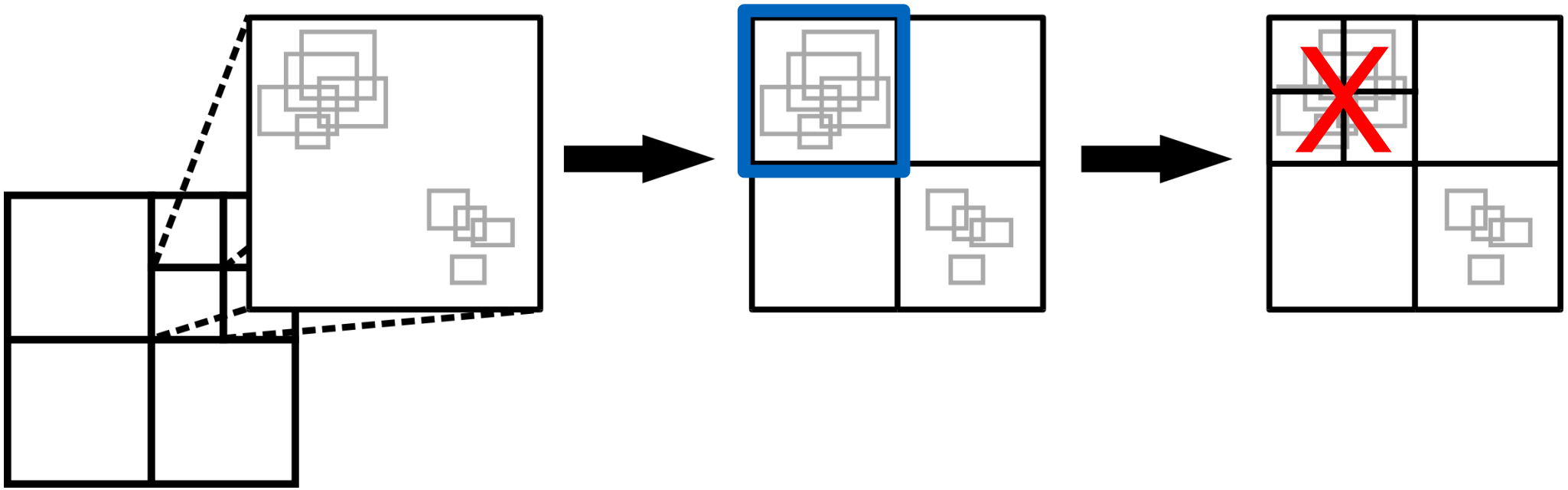
Query Load Balancing via Partitioning



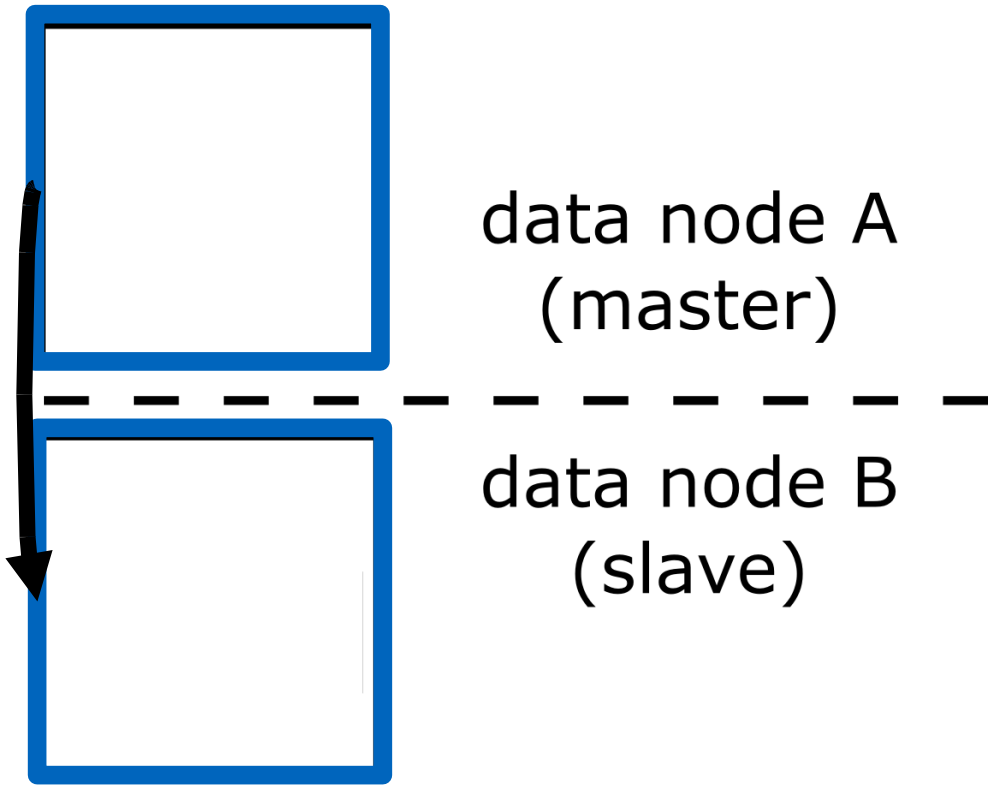
Query Load Balancing via Partitioning



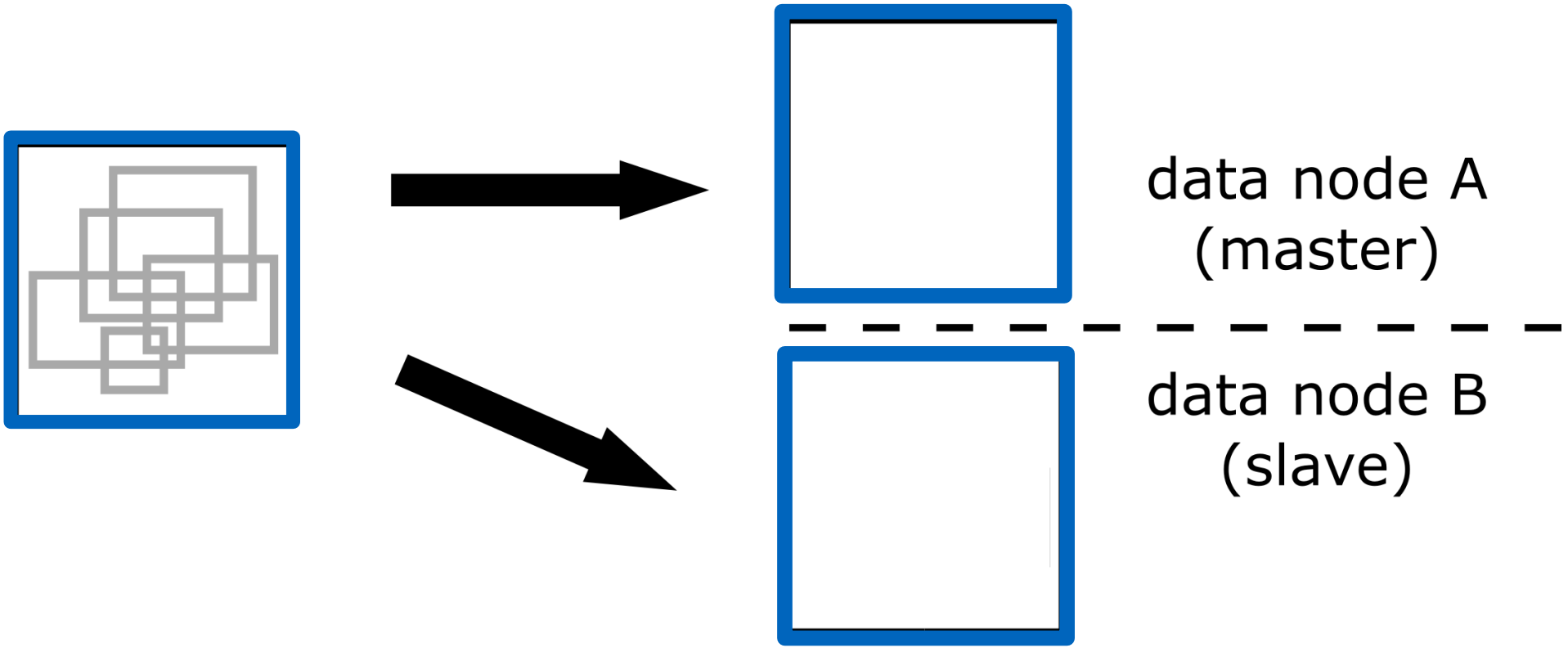
Query Load Balancing via Partitioning



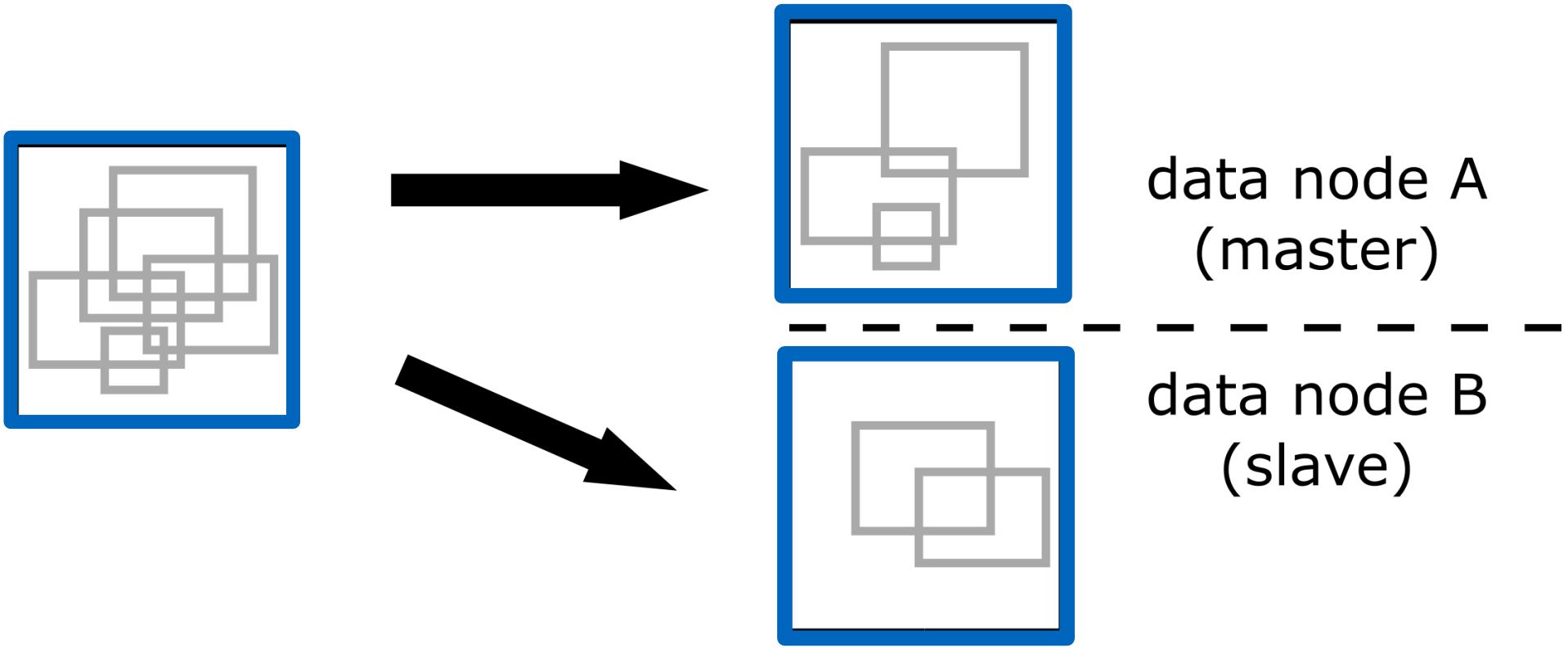
Query Load Balancing via Replication



Query Load Balancing via Replication

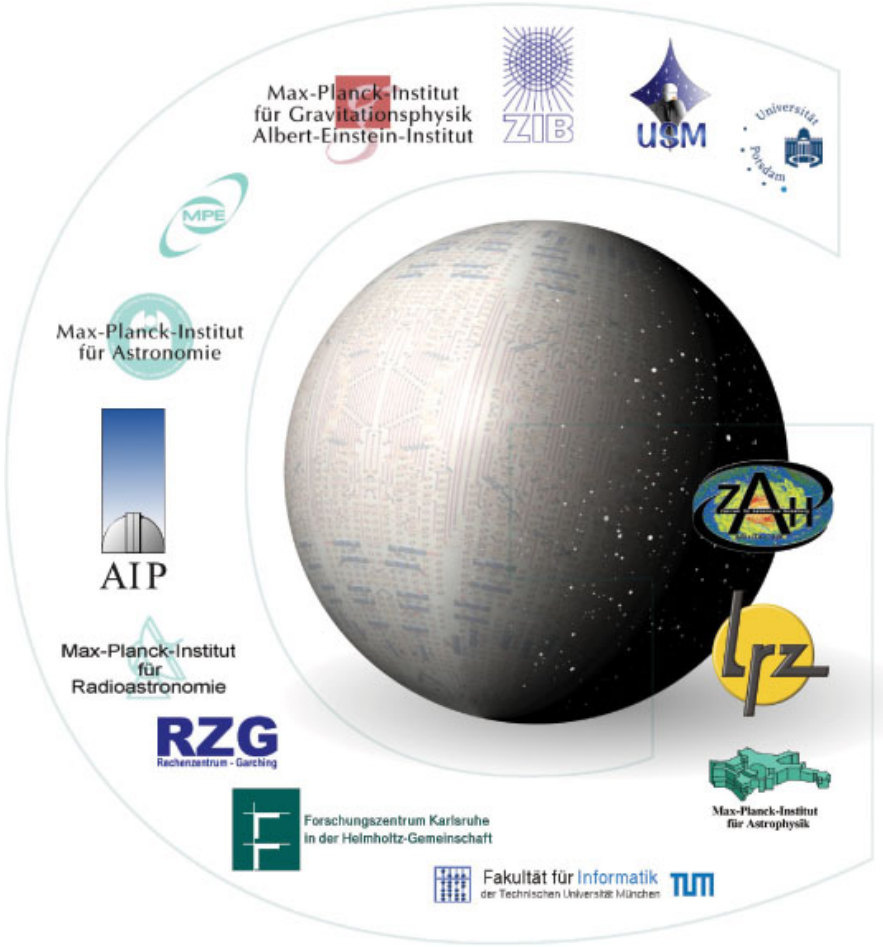


Query Load Balancing via Replication



The AstroGrid-D Project

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

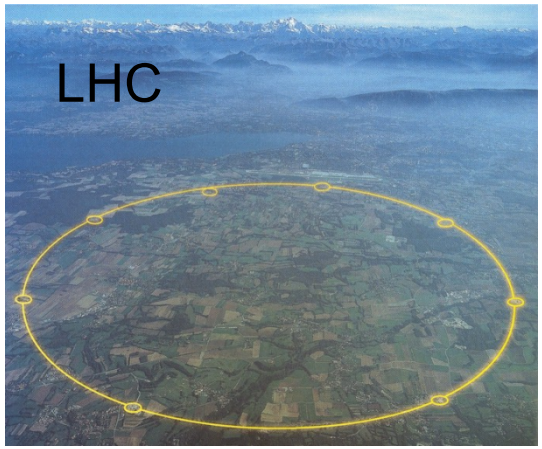


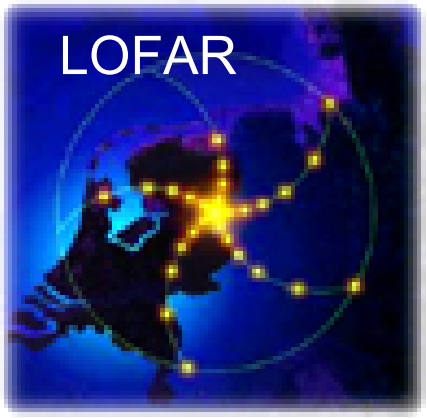
Federal Ministry of Education and Research

D-GRID

Up-Coming Data-Intensive Applications

- Alex Szalay, Jim Gray (Nature, 2006): "Science in an exponential world"
- Data rates
  - Terabytes a day/night
  - Petabytes a year
- LHC
- LSST
- LOFAR
- Pan-STARRS

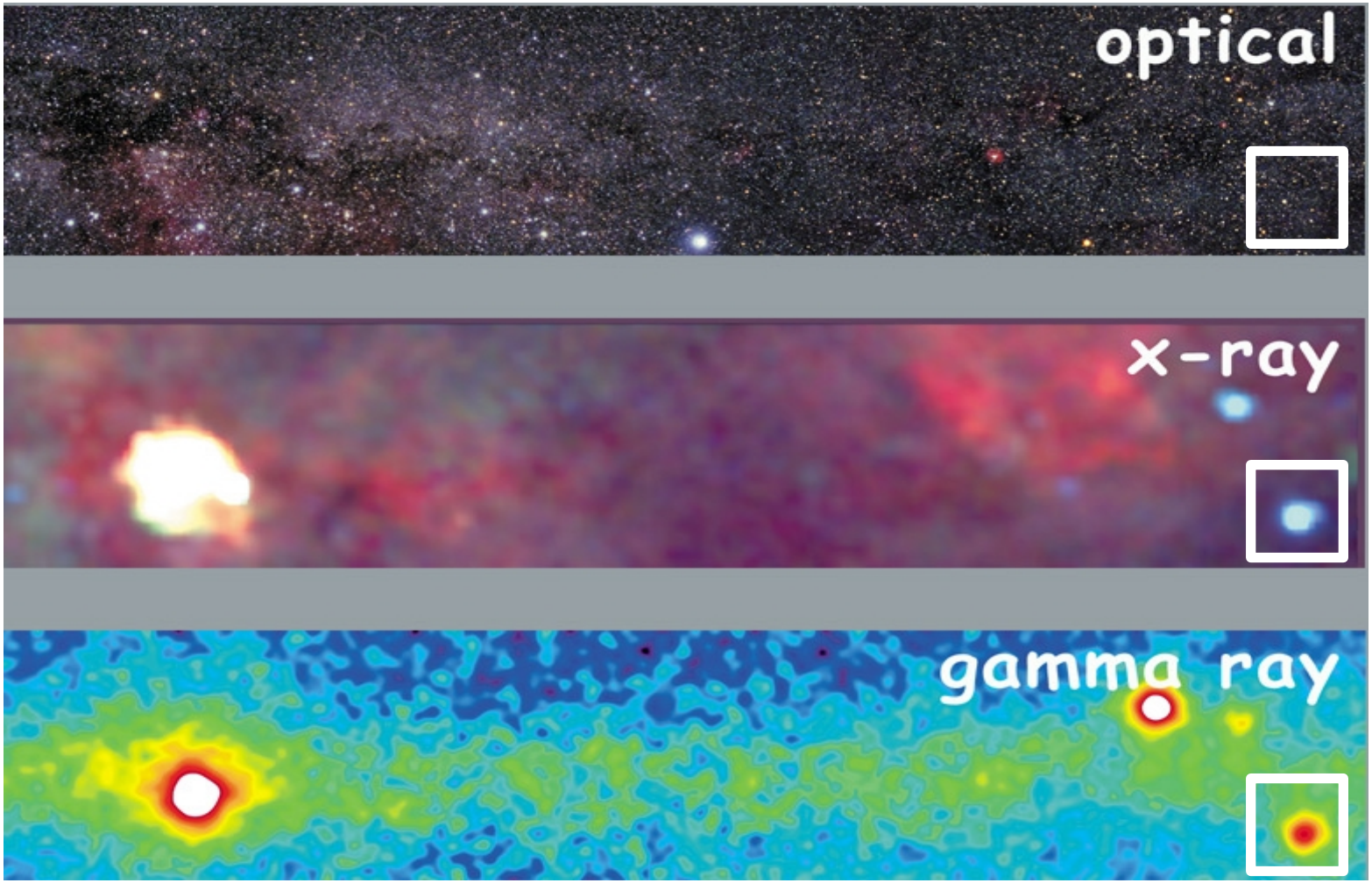




Pan-STARRS

Large Synoptic Survey Telescope

The Multiwavelength Milky Way



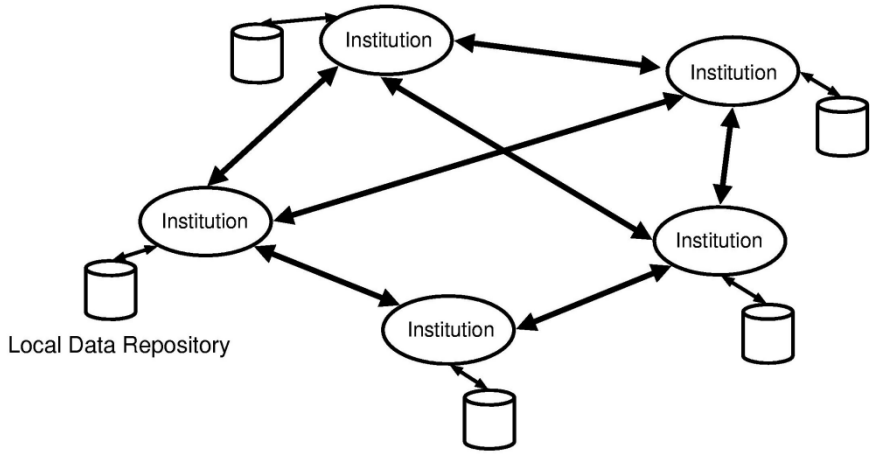
http://adc.gsfc.nasa.gov/mw/

Research Challenges

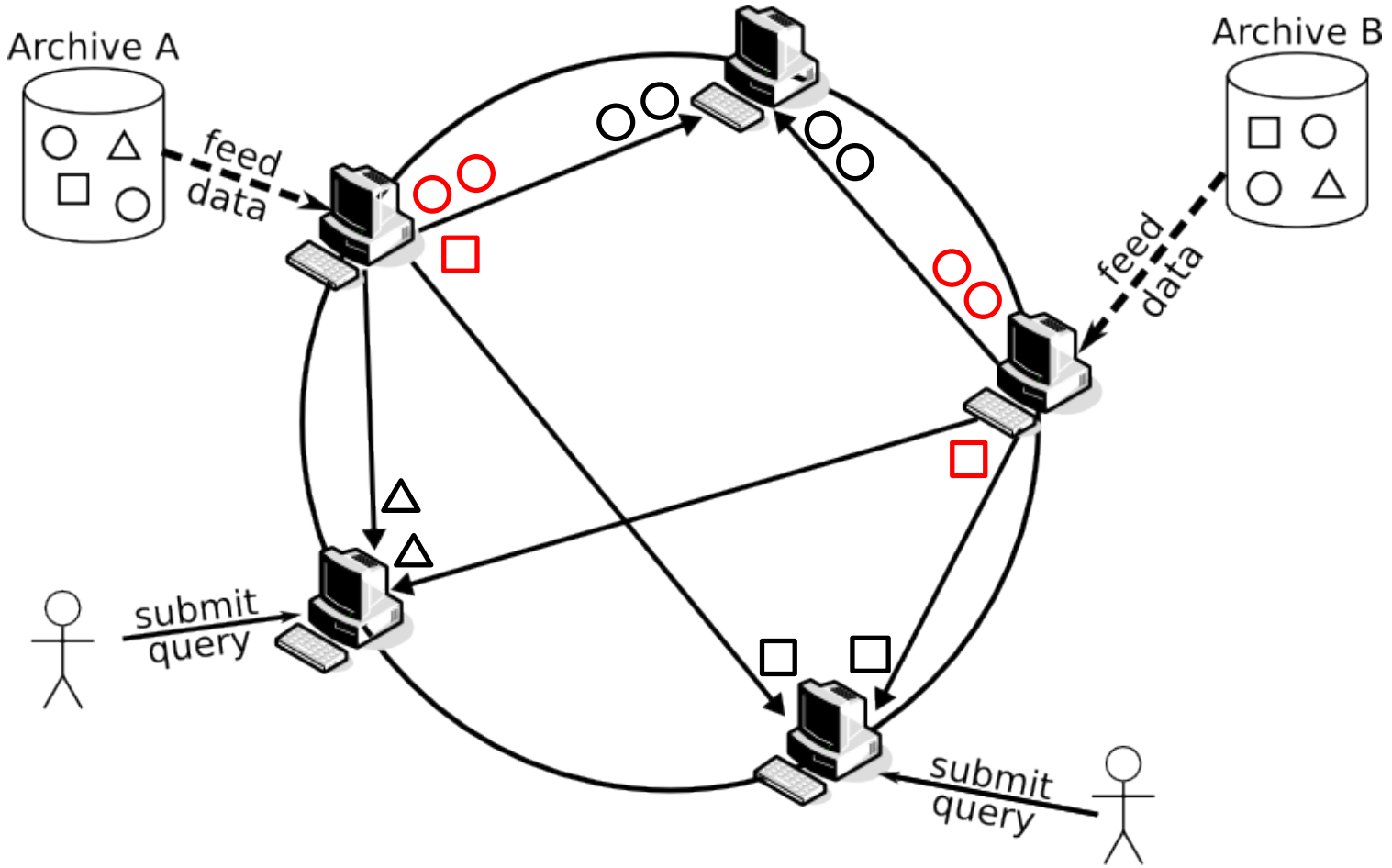
- Directly deal with Terabyte/Petabyte-scale data sets
- Integrate with existing community infrastructures
- High throughput for growing user communities

Current Sharing in Data Grids

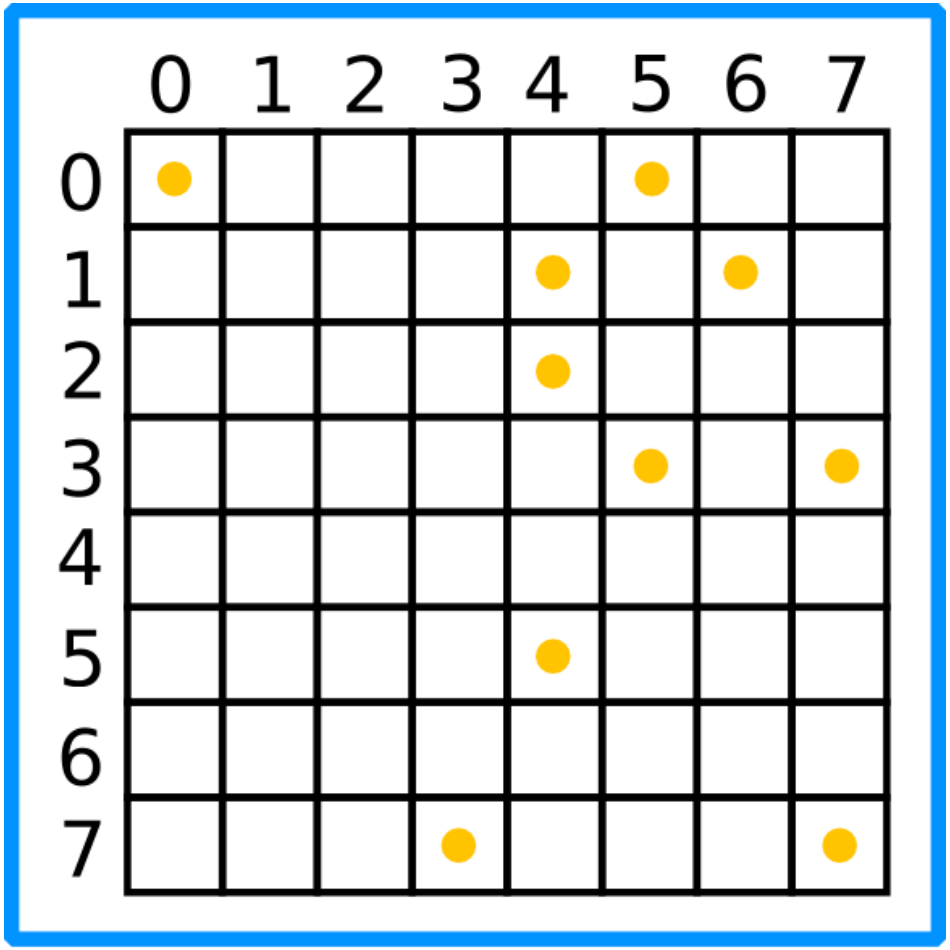
- Data autonomy
- Policies allow partners to access data
- Each institution ensures
  - Availability (replication)
  - Scalability
- Various organizational structures [Venugopal et al. 2006]:
  - Centralized
  - Hierarchical
  - Federated
  - Hybrid



Community-Driven Data Grids (HiSbase)

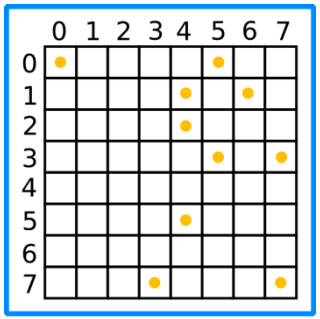


“Distribute by Region – not by Archive!”

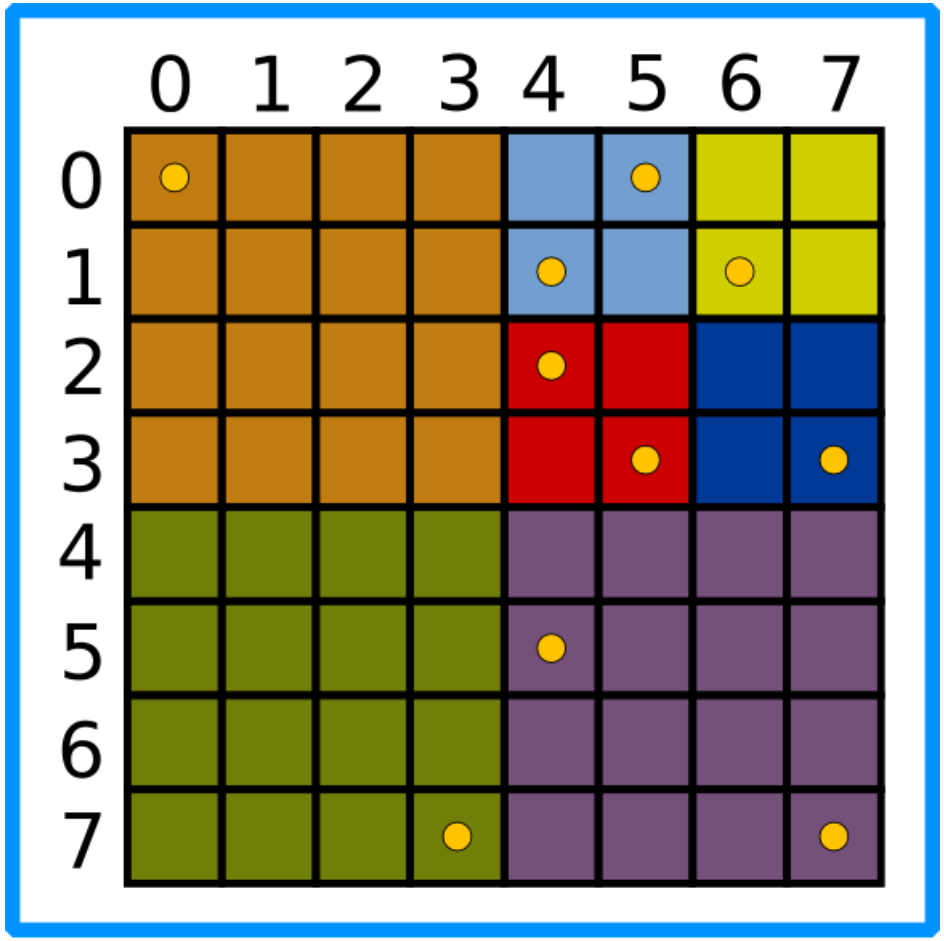


Training set

“Distribute by Region – not by Archive!”

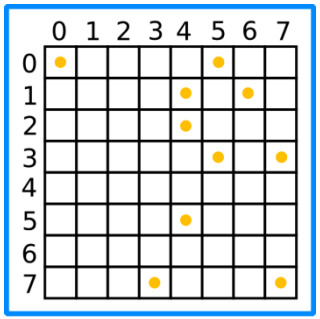


Training set

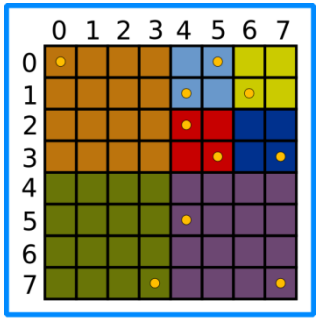


Histogram regions

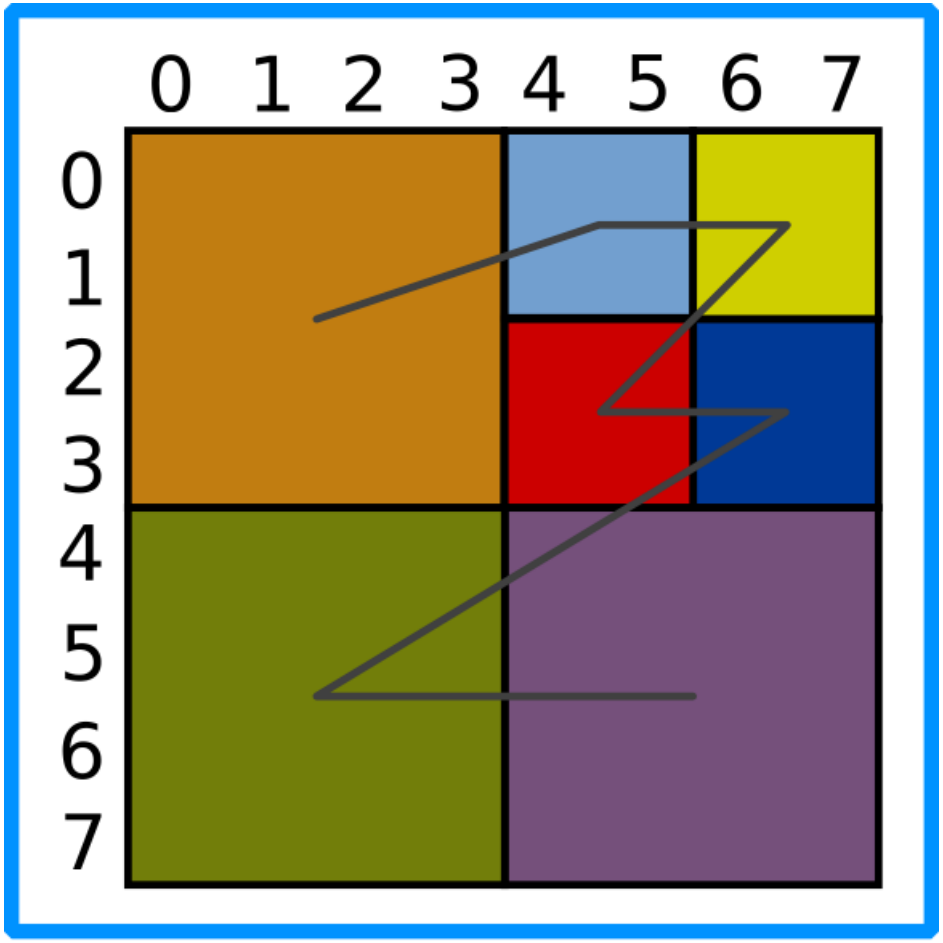
“Distribute by Region – not by Archive!”



Training set

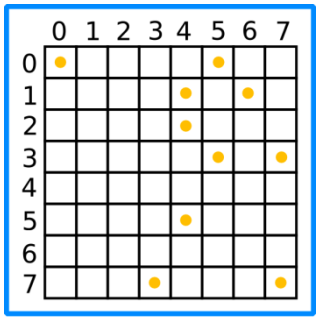


Histogram regions

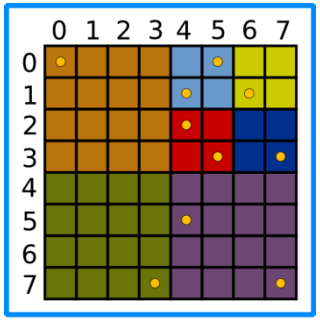


Z-Linearization

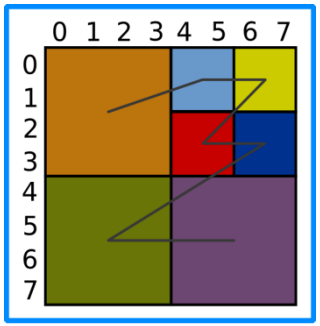
“Distribute by Region – not by Archive!”



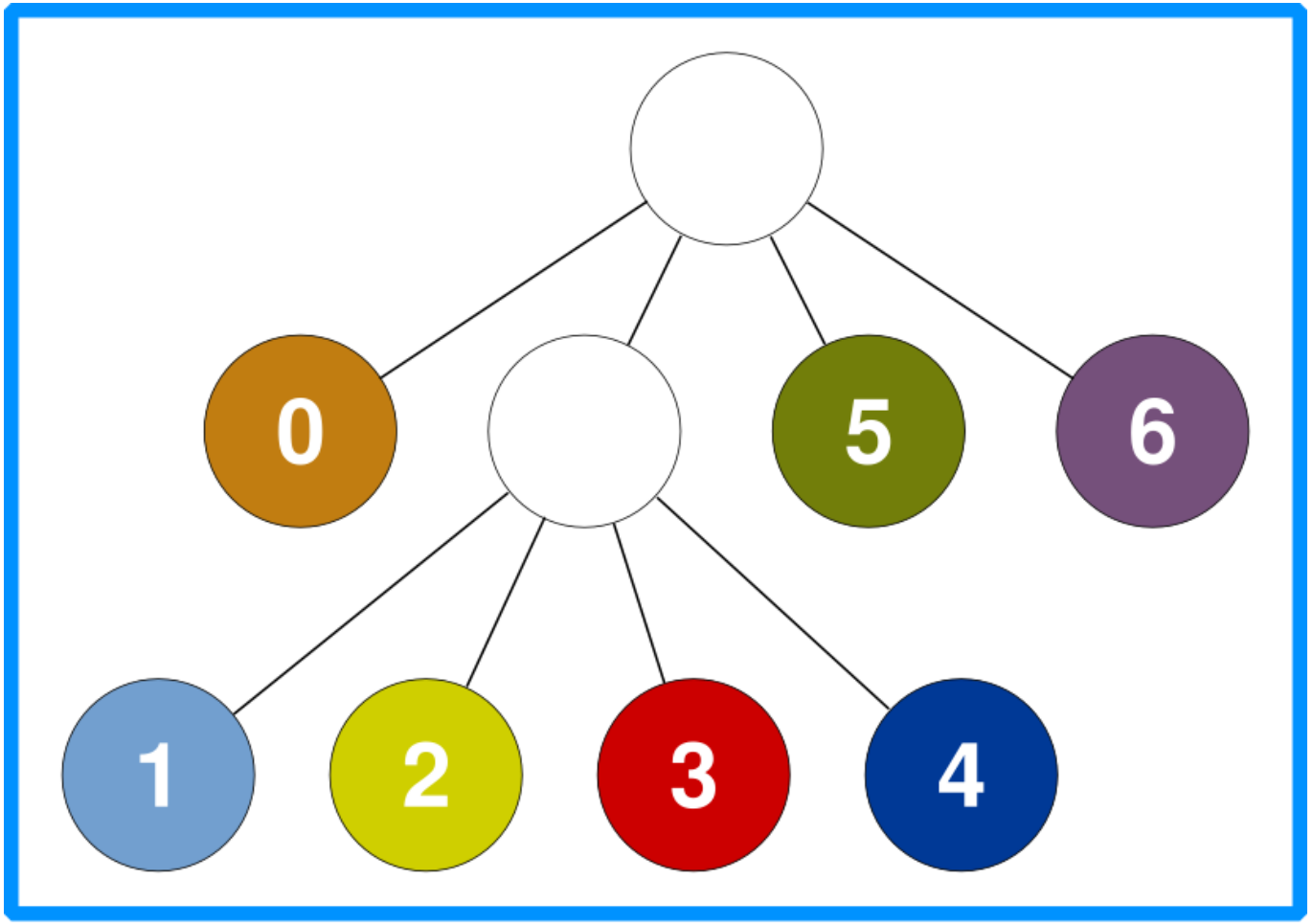
Training set



Histogram regions

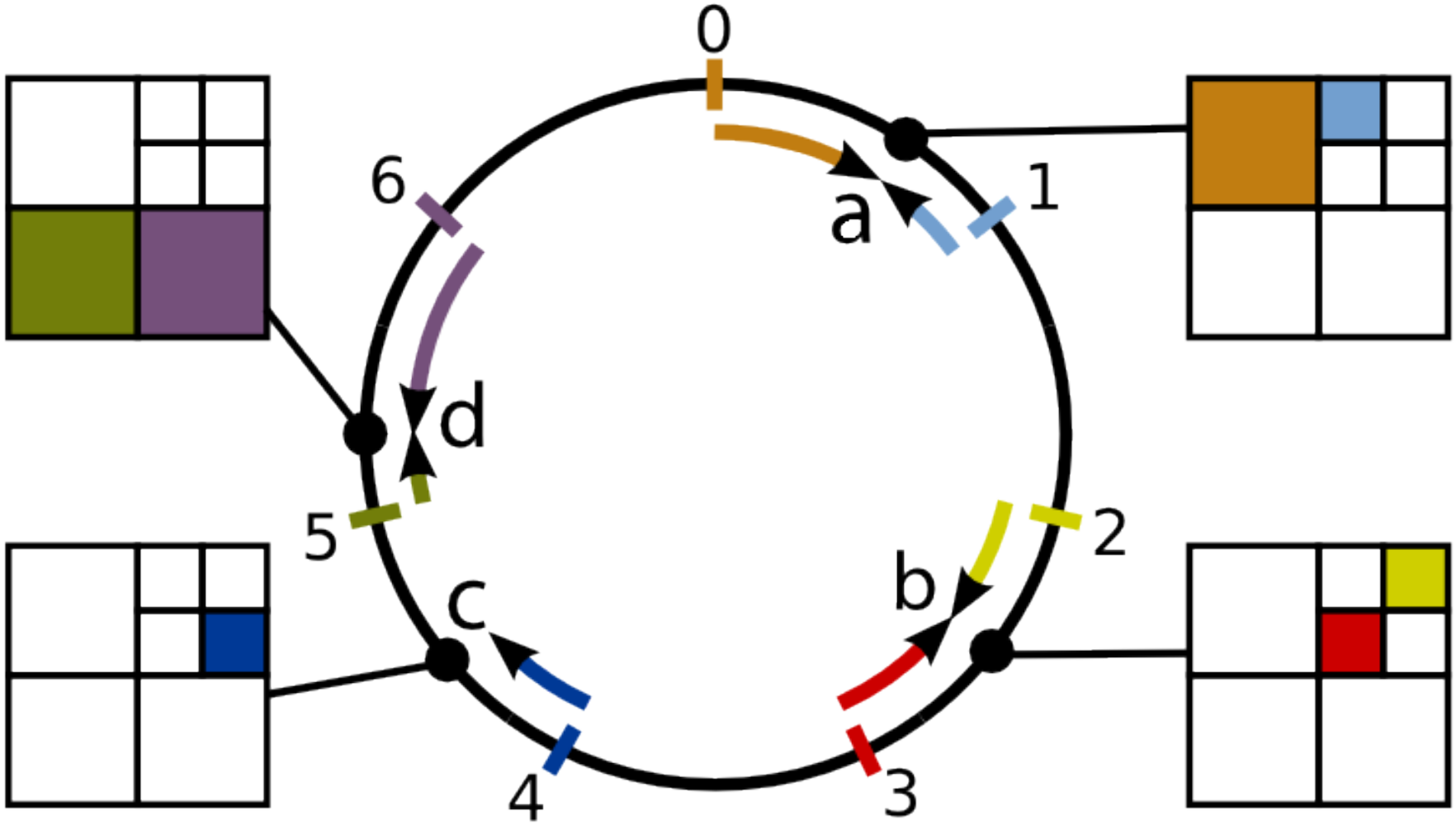


Z-Linearization



Quadtree

Mapping Data to Nodes

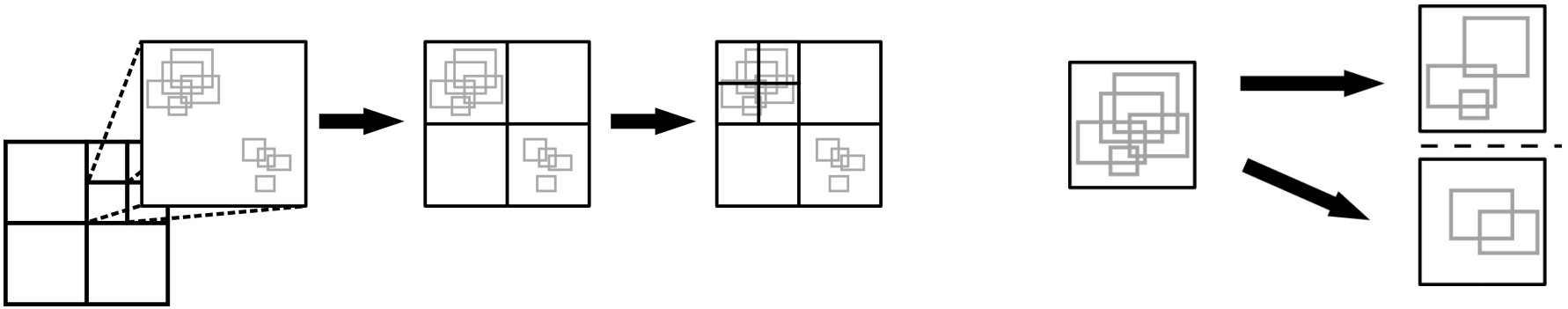


Workload-Aware Training Phase

- Incorporate query traces during training phase
- Base partitioning scheme on
  - Data load
  - Query load
- Challenges
  - Balance query load without losing data load balancing
  - Approximate real query hot spots from query sample

Dealing with Query Hot Spots

- Query skew triggered by increased interest in particular subsets of the data
- Two well-known query load balancing techniques:
  - Data partitioning
  - Data replication
- Finding trade-offs between both



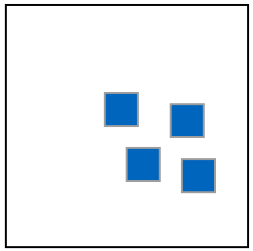
When to Split (Partition) or to Replicate

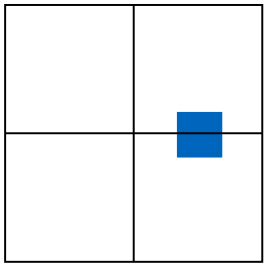
- Considers partition characteristics
  - Amount of data (few/many data points)
  - Number of queries (few/many queries)
  - Extent of regions and queries (small/big queries)

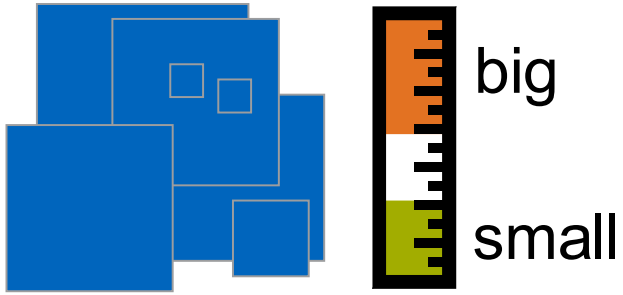
Data points	Few Queries		Many Queries	
	Small	Big	Small	Big
Few	–	–	SPLIT	REPLICATE
Many	SPLIT	SPLIT	SPLIT	REPLICATE

Region Weight Functions

- Data only (#objects in a region)
- Queries only (#queries in a region)
- Scaled queries
  - Approximate “real” extent of hot spot
  - Avoid overfitting to training query set
- Heat of a region (#objects * #queries)
- Extents of regions and queries
  - Replicate when many big queries

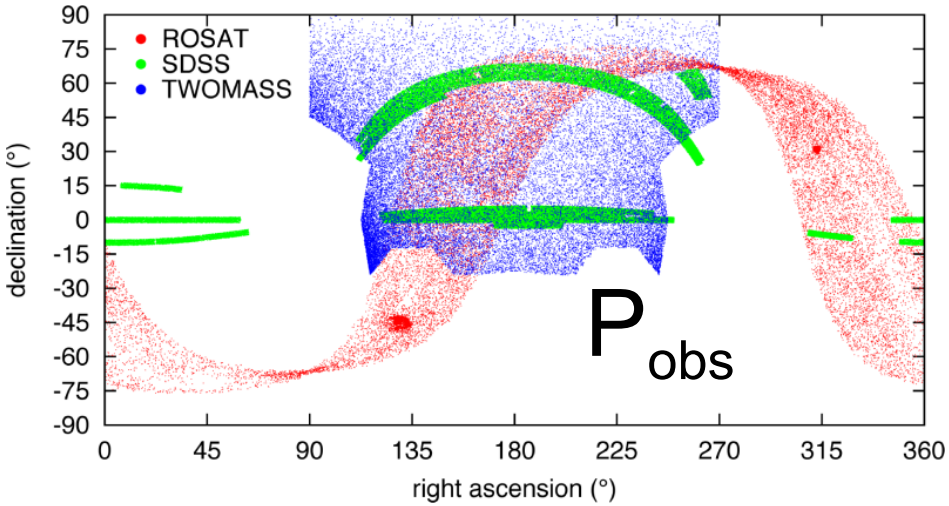






Evaluation

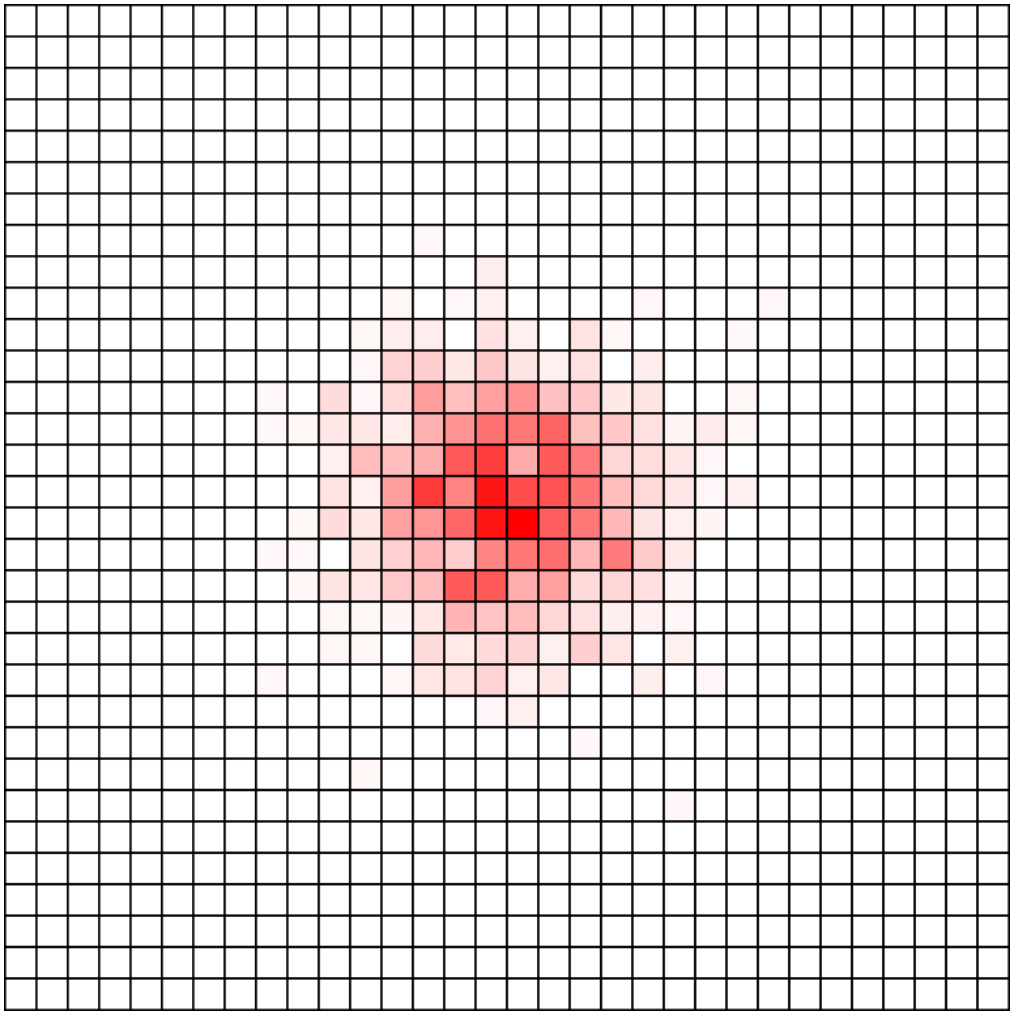
- Weight functions: data, heat, extent
- Data sets (observational, simulation)
- Workloads (SDSS query log, synthetic)
- Partitioning Scheme Properties
  - Load distribution
  - Communication overhead
- Throughput Measurements
  - Distributed setup
  - FreePastry simulator



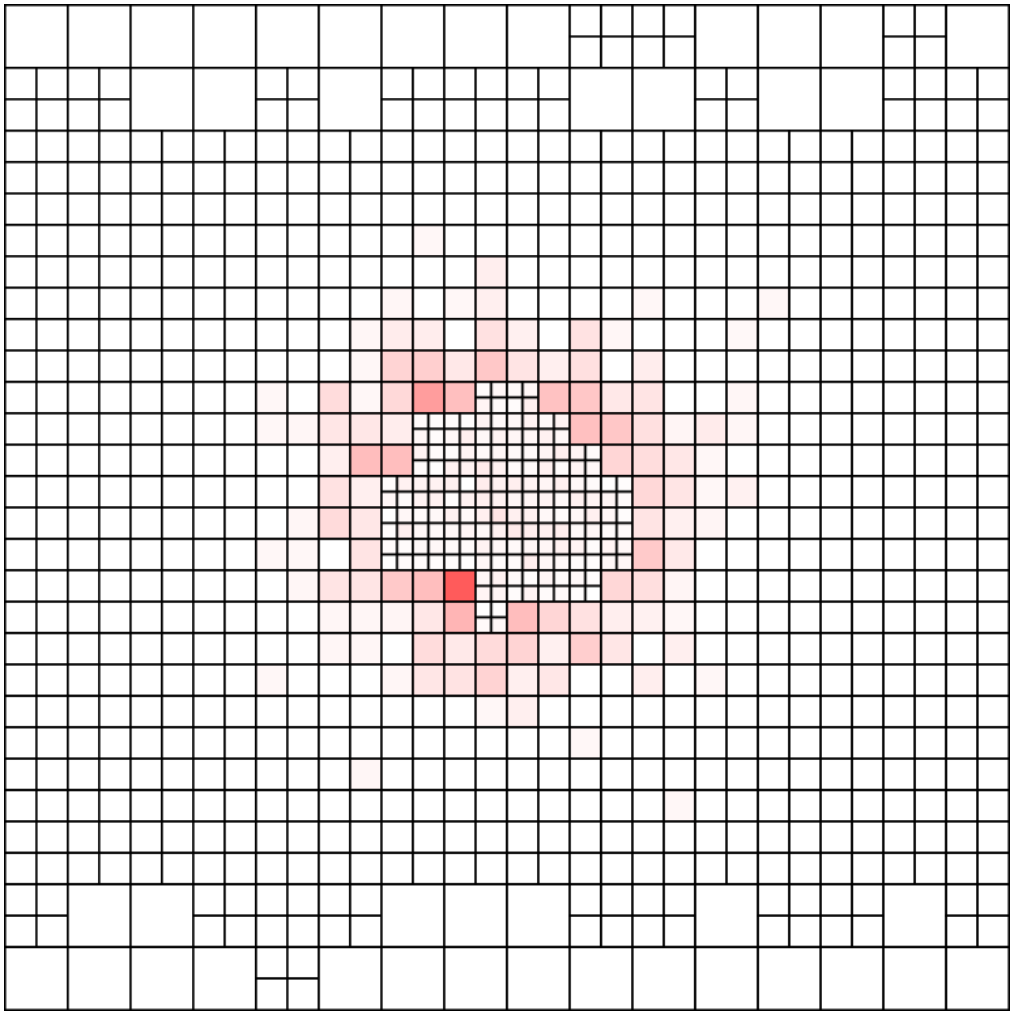
Load Distribution

- Uniform data set from the Millennium simulation
- Workload with extreme hot spot
- In the following:
  - 1024 partitions
  - Heat of a region (#data * #queries)
  - Normalized across all partitioning schemes

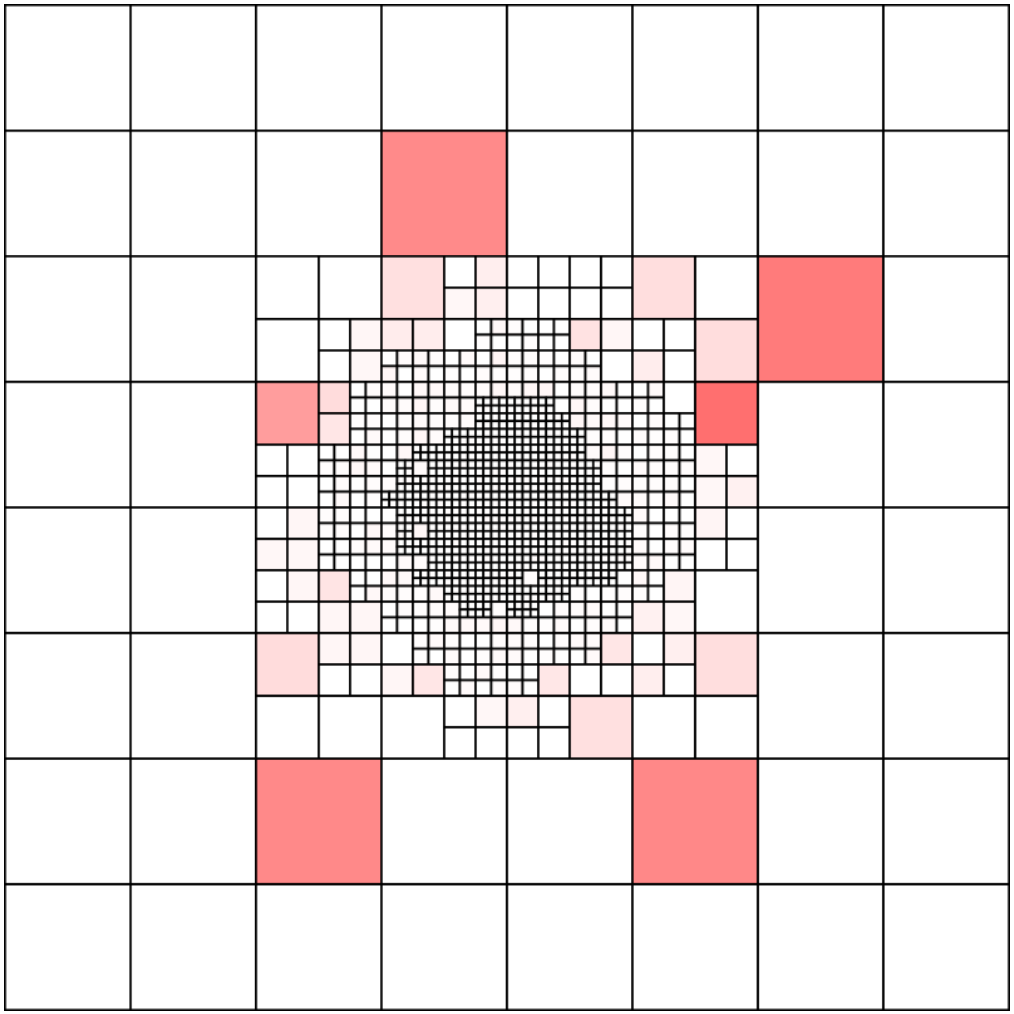
Query-unaware Training



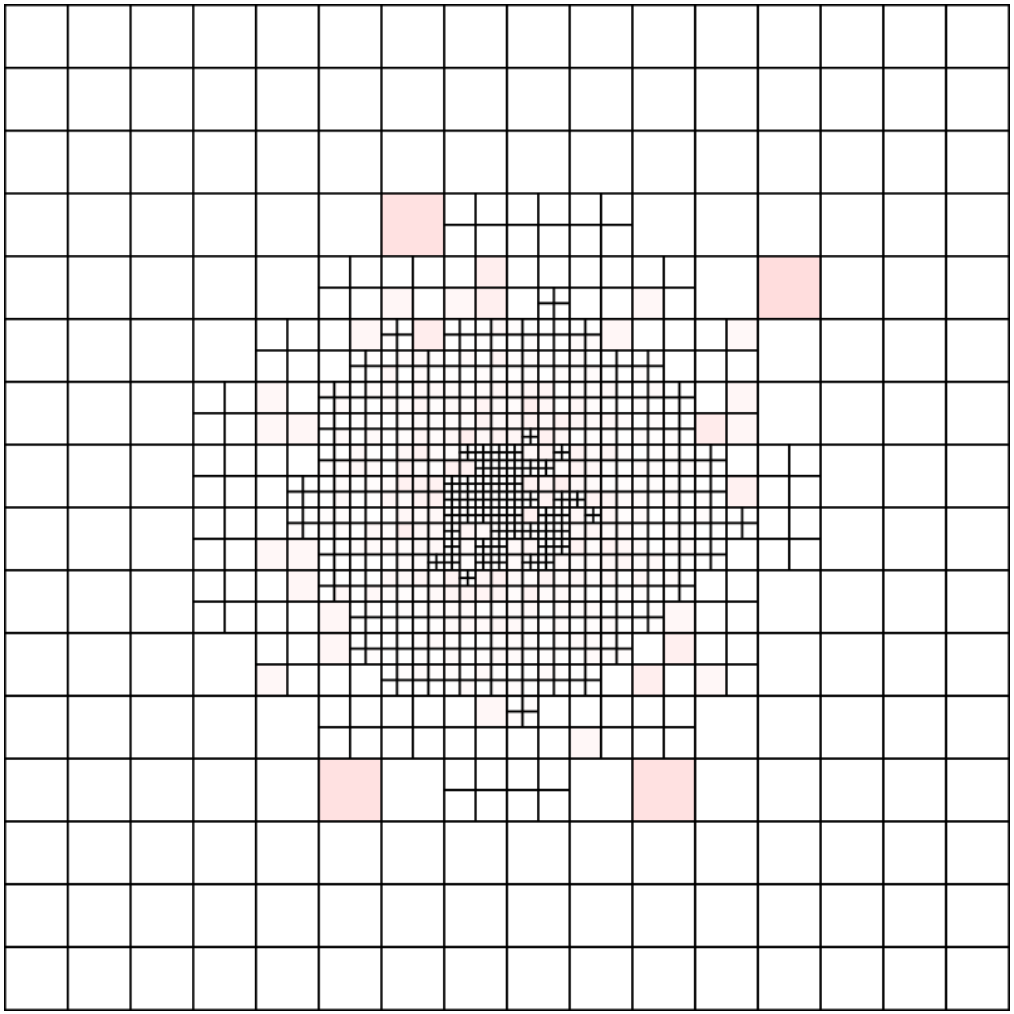
Training with Scaled Queries (scaled 50x)



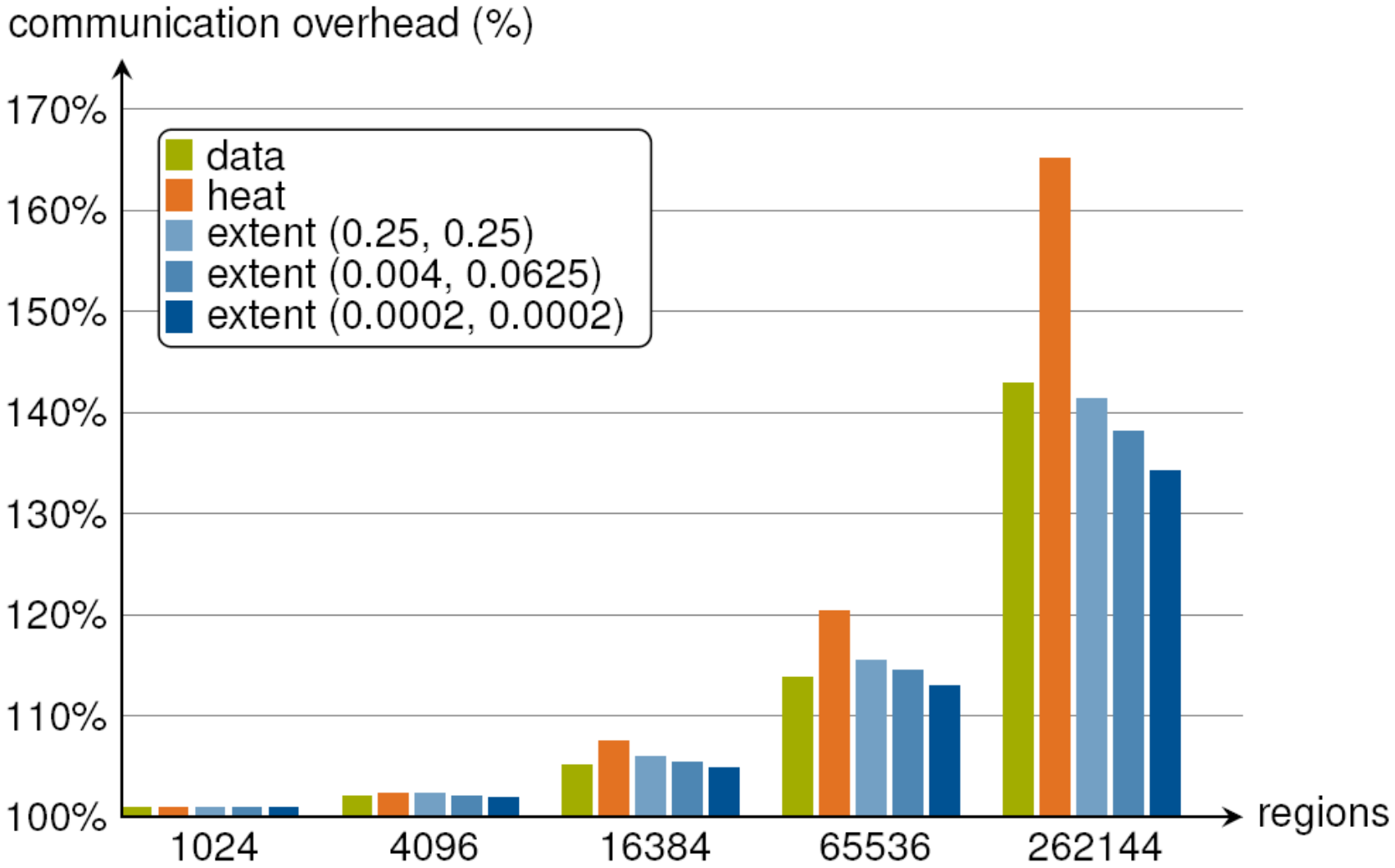
Training with Scaled Queries (scaled 400x)



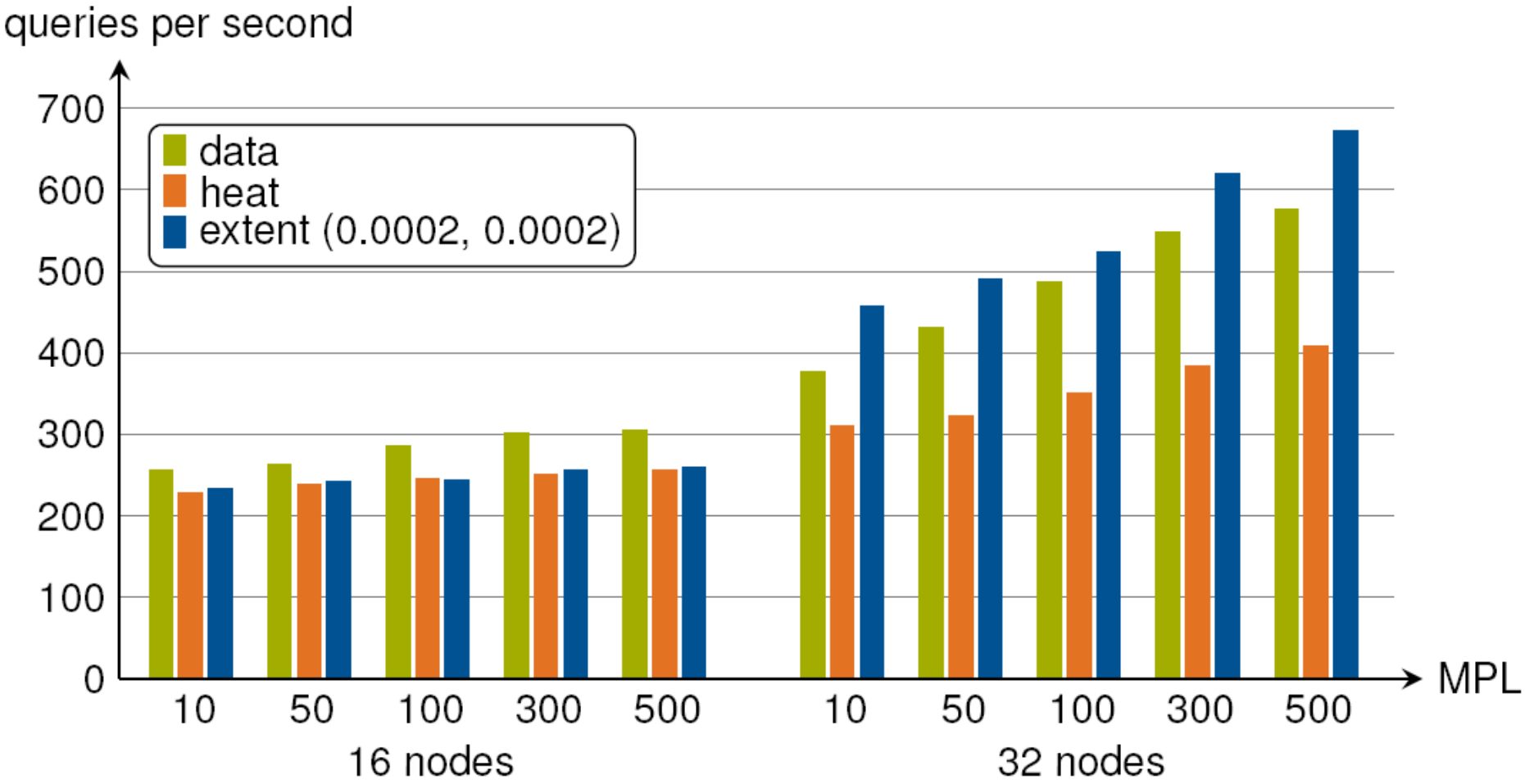
Heat-based, Extent-based Training



Communication Overhead for P_{obs}



Throughput for P_{obs}

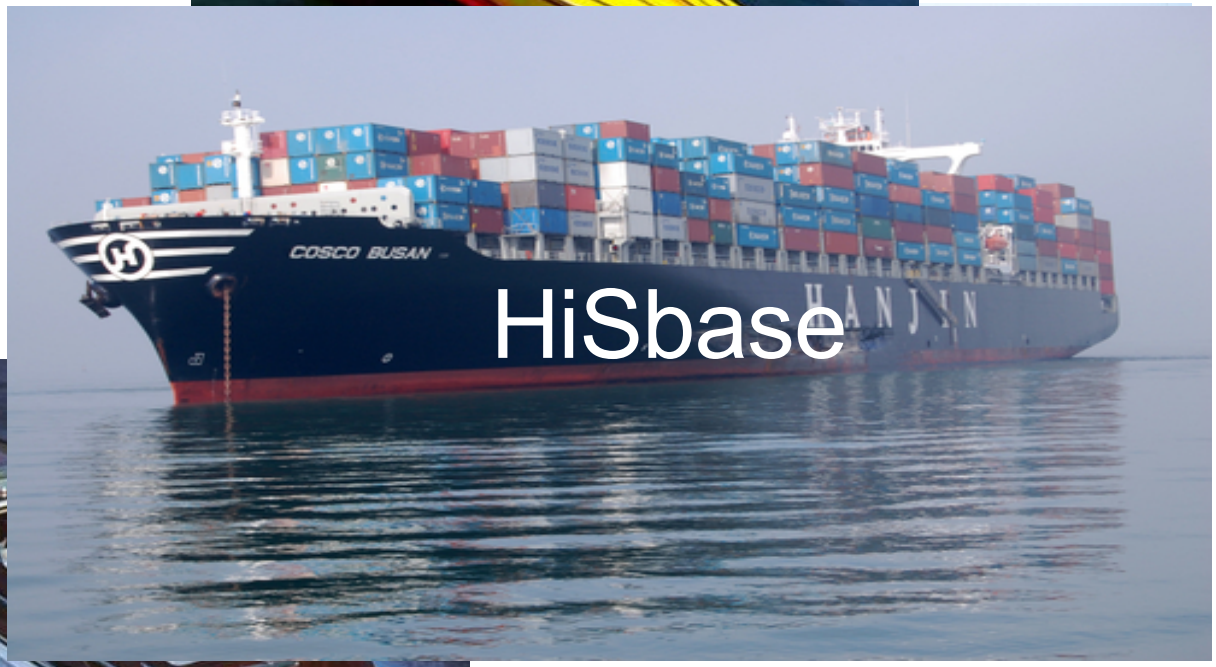


Load Balancing During Runtime

- Complement workload-aware partitioning with runtime load-balancing
- Short-term peaks
  - Master-slave approach
  - Load monitoring
- Long-term trends
  - Based on load monitoring
  - Histogram evolution

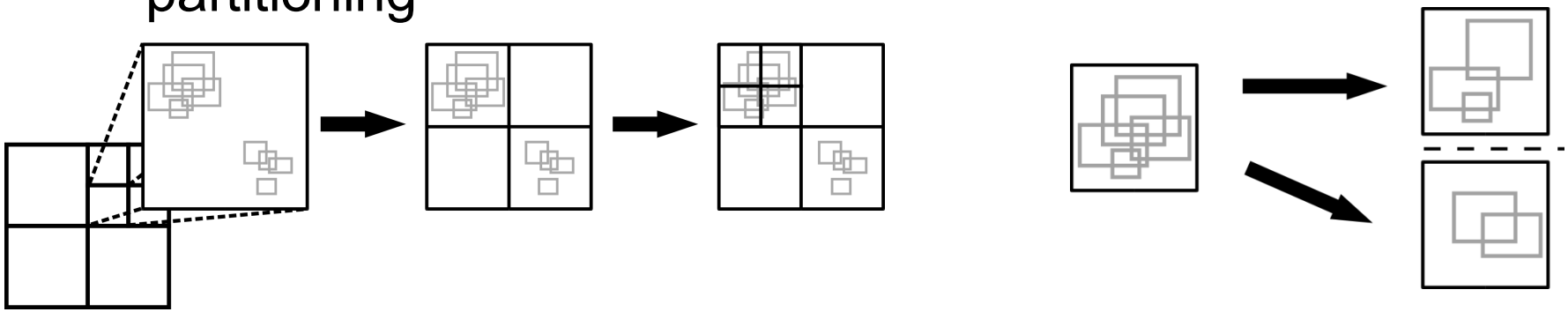
Related Work

- On-line load balancing
- Hundreds of thousands to millions of nodes
- Reacting fast
- Treating objects individually



Should I Split or Replicate?

- Many challenges and opportunities in e-science for database research
  - High-throughput data management
  - Correlation of distributed data sources
- Community-driven data grids
  - Dealing with data skew and query hot spots
  - Workload-awareness by employing cost model during partitioning

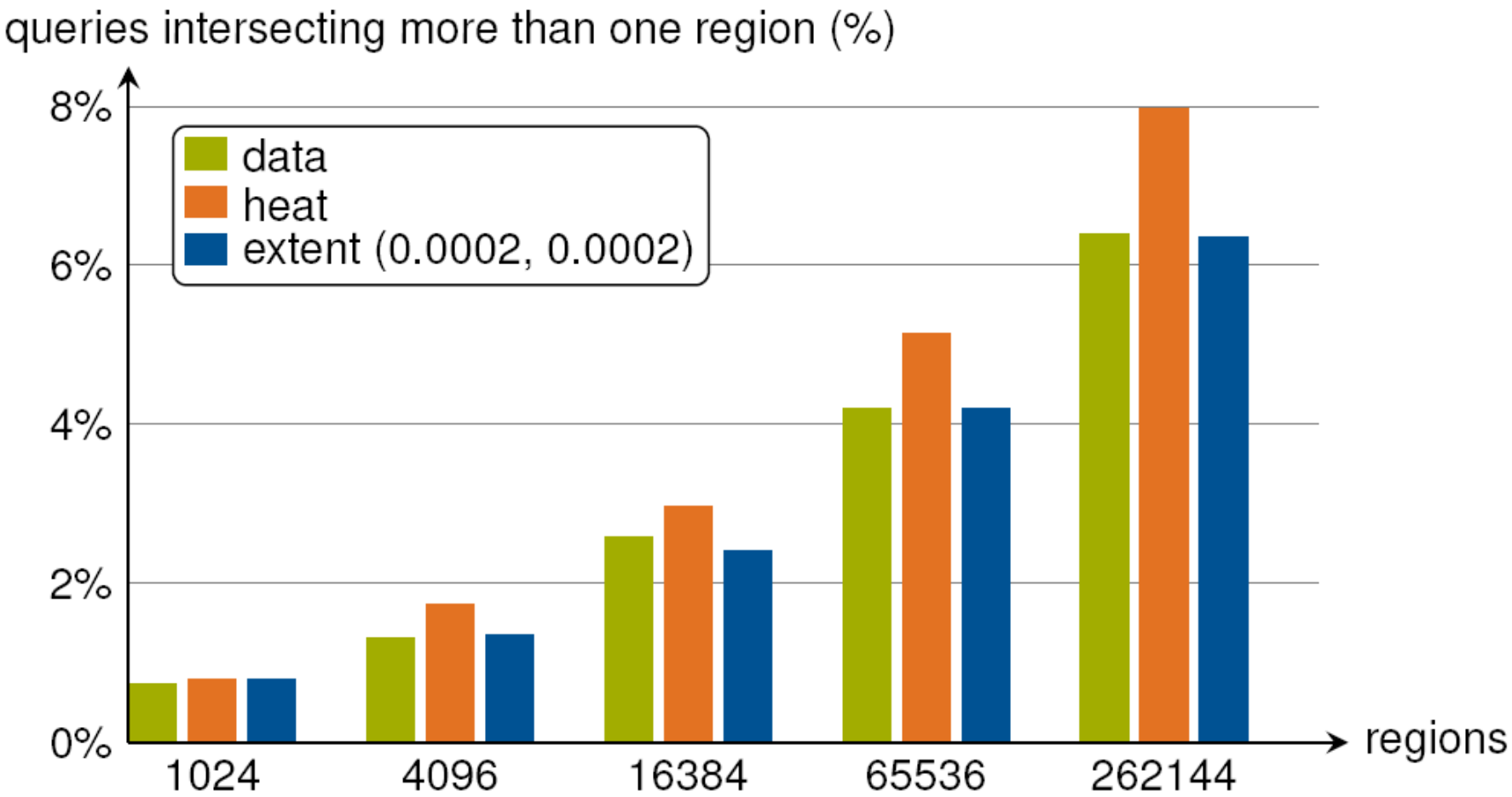


Get in Touch

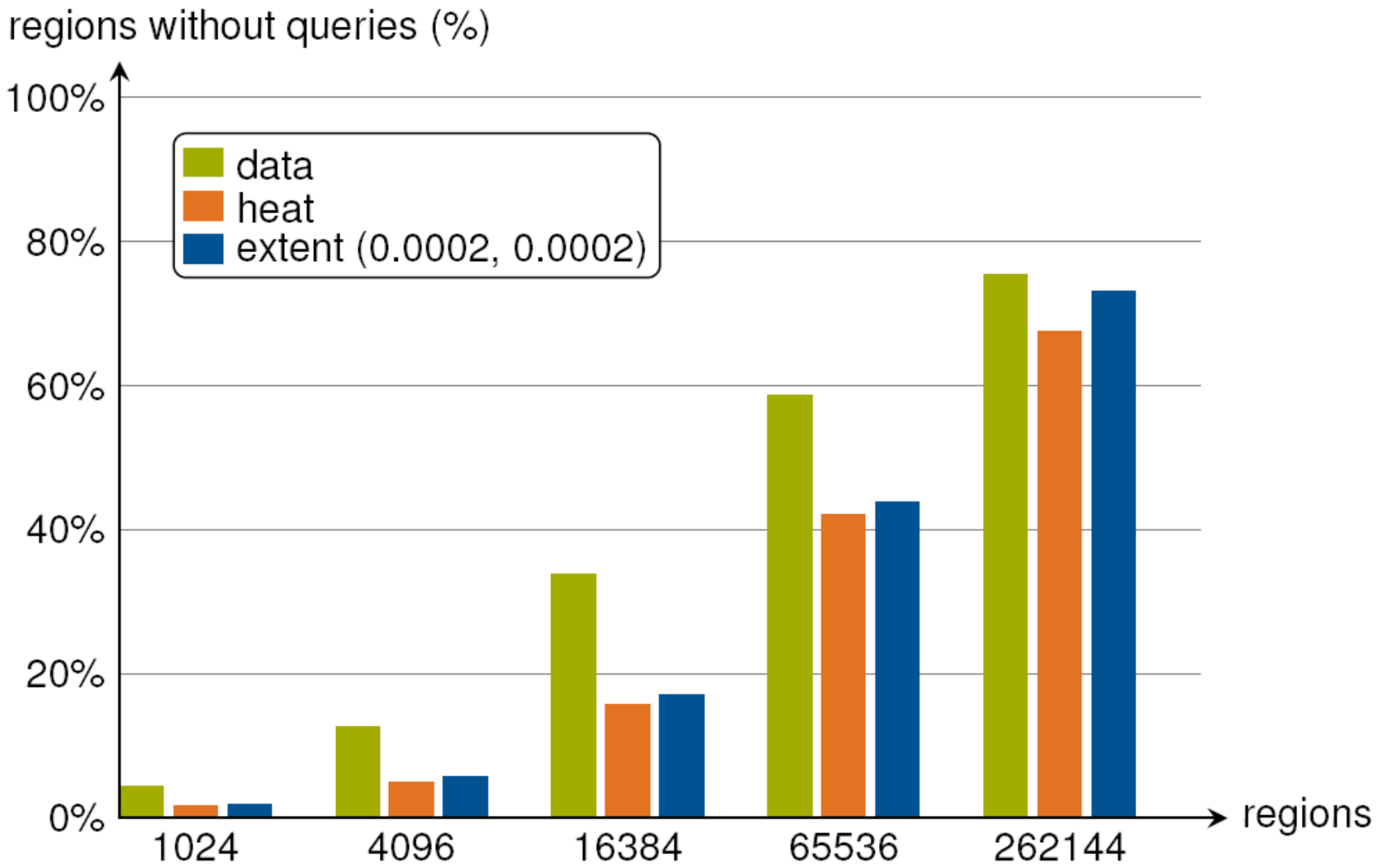
- Database systems group, TU München
  - Web site: http://www-db.in.tum.de
  - E-mail: scholl@in.tum.de
- The HiSbase project
  - http://www-db.in.tum.de/research/projects/hisbase/

Thank You for Your Attention

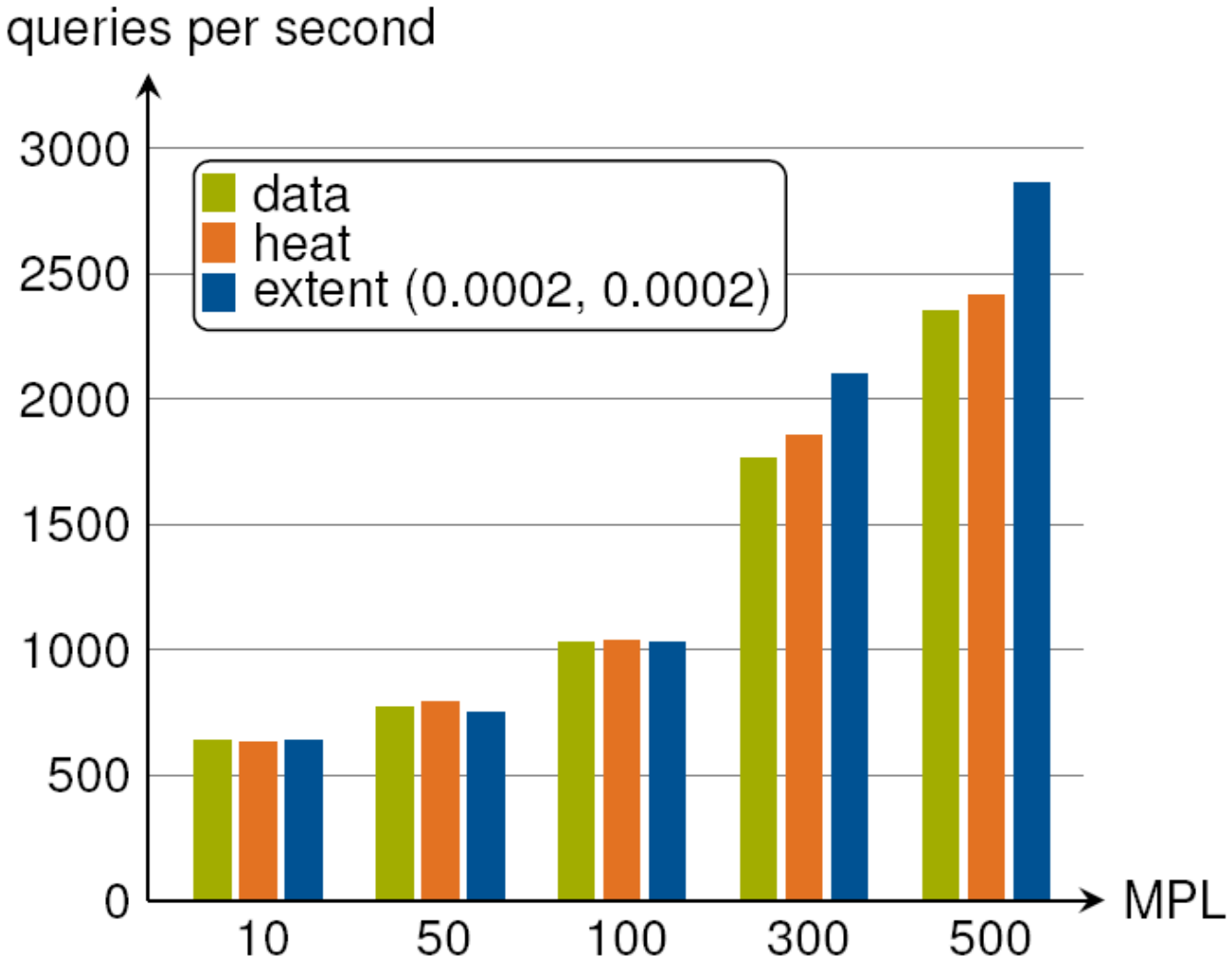
Queries Intersecting Multiple Regions



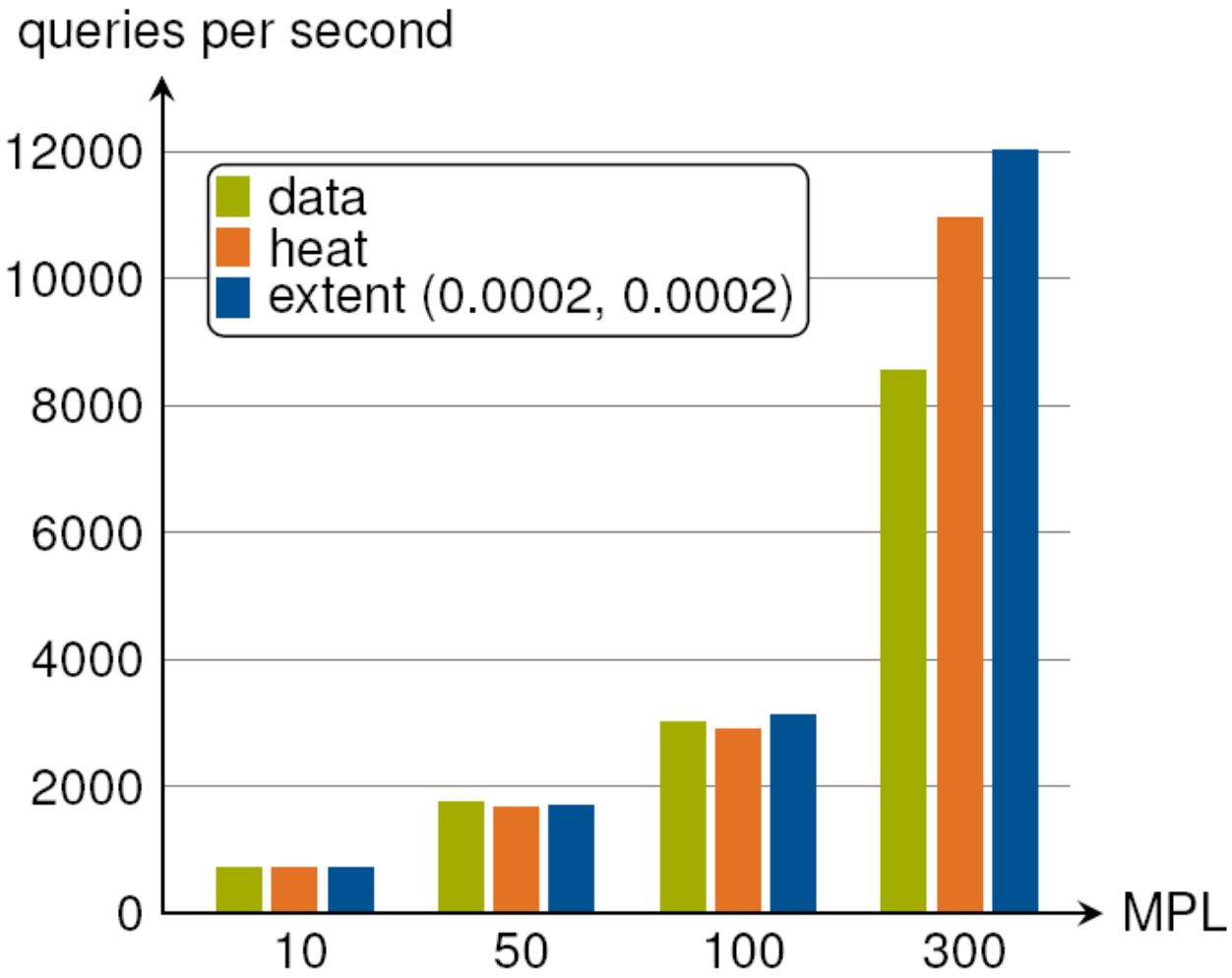
Regions Without Queries



Throughput for P_{obs} (300 nodes, sim.)



Throughput for P_{obs} (1000 nodes, sim.)



Throughput (Region-Uniform Queries)