Open SkyQuery, SkyNodes, and ADQL

Status and Future Plans

Maria A. Nieto-Santisteban
Johns Hopkins University

Issues with Open SkyQuery

People (want to) use OSQ but at the same time the 5k row limitation discourage them

- Why the 5K?
  - OSQ and SkyNodes are synchronous
  - Inefficient transfer format between nodes (VOTable)
  - Lack of mechanisms to manage results (VOSpace)
- We are working on solving the 5k limit with mid and long term plans
- ADQL issues
  - The XML representation is not helping
  - SkyNodes lack methods using ADQL/s directly

OSQ and SkyNode Development

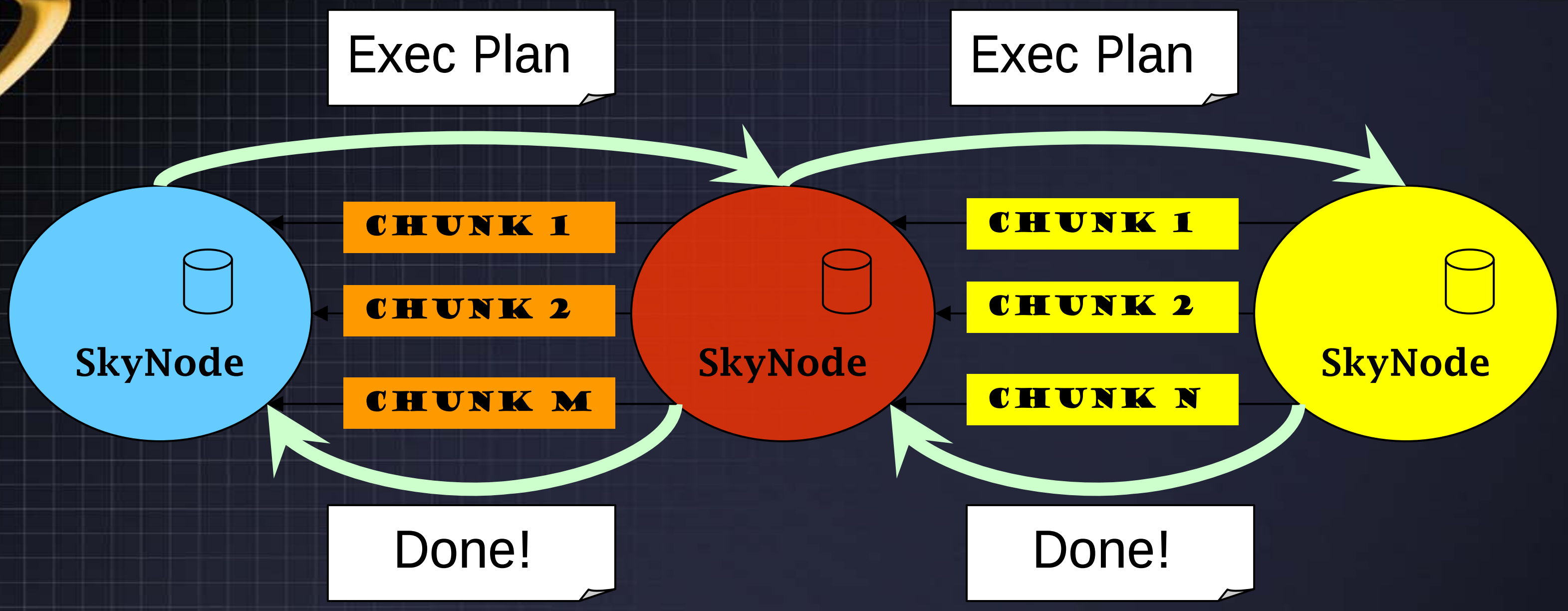
- Dev.openskyquery.net
  - Implements ADQL 0.91
  - Solves many know bugs
  - Displays ROI using SIAP services
  - Portal still limited to 5k
  - SkyNodes can change the limit for individual access

- Why is not dev.openskyquery.net “live”?
  - ADQL/SkyNode 1.0 not finished yet
  - SkyNodes not hosted at JHU would fail

Two-Steps Development Plan

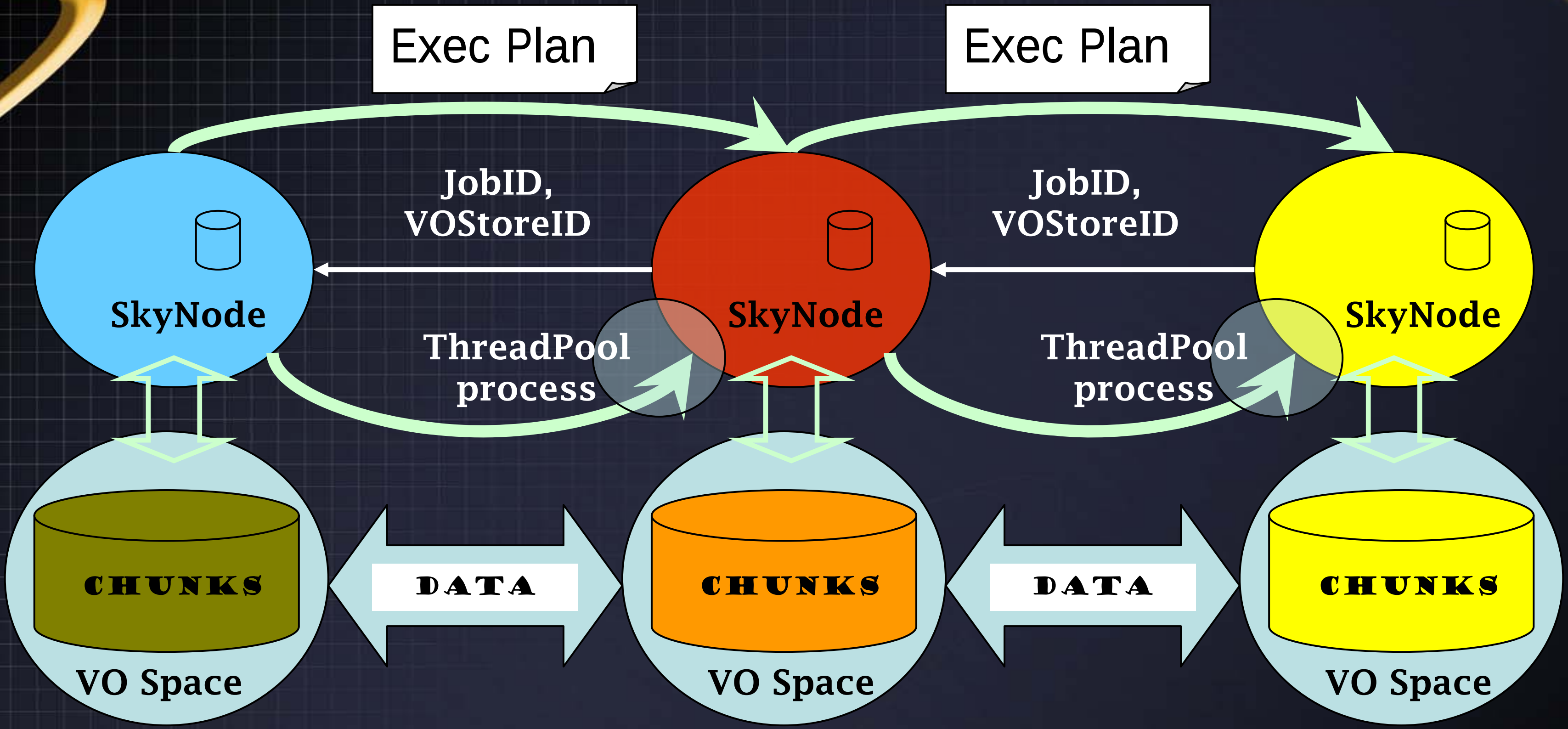
- Mid term -> bring 5k to 50k
  - Add pipeline processing in the current implementation
  - Optimize queries at the DB level
  - Implement XMATCH using zones instead of HTM
- Long term -> Unlimited rows
  - Put results into VOSpace
  - Add Asynchronism
  - Add Authentication

Mid term SkyNode design



- SkyNodes split results in chunks
- Chunks are processed by the next node as soon as they are available
- It requires a queue/tracking mechanism checking no data has been lost
- Working in preliminary prototype stage
- It allows already queries that return about 50K rows

Long term SkyNode design



- The next step (unlimited size problem) needs:
  - A (VO)Space to put results. Mechanisms for fast transfer and space management
  - Asynchronous mechanisms that allow long time query execution
  - Authentication procedures to let users access their results

Improvements at the DBMS level

- Optimizing schema and queries
- Zones and Parallelism
  - ZoneID = floor (dec + 90.0 / ZoneHeight)
  - Improves in many (most) cases the performance of neighborhood searches
  - It is based on relational algebra only

    => full independency from HTM libraries
  - Provides a simple base for partitioning and parallelism
  - XMATCH can be efficiently processed in parallel using zones

ADQL issues @ the portal

- The ADQL CORE + Extensions design doesn’t quite work for OSQ (as a federation of databases)
- OSQ will support
  - Single Node – Multiple Tables (including REGION constraints)
  - XMATCH Queries (multiple nodes)
- OSQ cannot support Multiple Node – Multiple Table queries
  - It is not feasible to do full table JOINs across federation of nodes. Such capabilities require environments as CasJobs

Future Plans

- Mid Term
  - Fix some of the loose ends in the development portal
  - Add the minimum changes to bring the 5k limit to 50k
  - Add zones to do XMATCH and neighbor searches

- Long Term
  - Multithreaded SkyNode
  - Threadpool procedures
  - Message Queue management and efficient data transfer
    - Format: MTOM, gzip, VOTable?
  - Footprint services to optimize XMATCH and Region queries
  - A (VO)Space to put results. Mechanisms for fast transfer and space management
  - Asynchronous mechanisms that allow long time query execution
  - Authentication procedures to let users access their results

- Ferriswheel – next generation
  - Putting in sync (federated) SkyNodes to optimize query processing