edikt

National e-Science Centre

EdSkyQuery-G Overview

Brian Hills, December 2004

www.edikt.org

Contents

- Edikt
- Motivation & Aims
- Architecture
- Current Status
- Results
- Future Outlook

Edikt

- **“e-Science Data, Information and Knowledge Transformation”**
  - Bridge the gap between ***applications*** and ***computer*** science:
    - Produce robust tools…
    - …for real application science problems…
    - …test them under extreme science conditions…
    - …and keep an eye on the commercial possibilities.
  - Projects which may be of interest to astronomers:
    - **BinX**, **Eldas** & **EdSkyquery-G.**
  - Visit: www.edikt.org

edikt

Astronomy Requirements

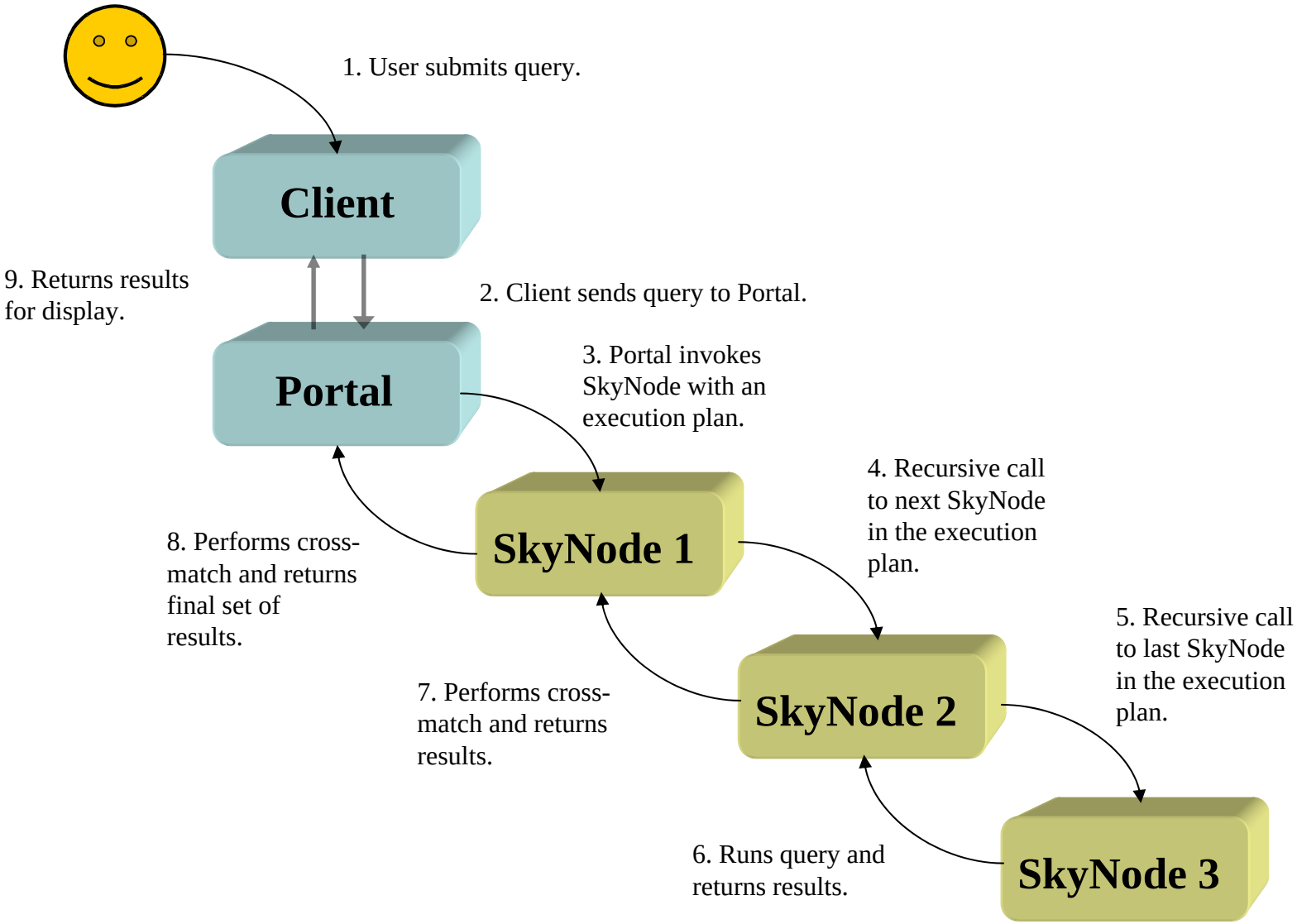
- Sky Surveys collect masses of data to be managed:
  - For example, Sloan Digital Sky Survey: 15TB.
  - 2 * 10GB databases to be used for the project.
  - Edikt will have access to a 155TB SAN.
- Further research by leveraging data from different surveys:
  - Must identify same object from different catalogues.
- Require a “federated” view:
  - Data is distributed, homogeneous, large scale.
  - Building one big data warehouse isn’t feasible.
  - Interoperable services to combine disparate data sources.

Is the middleware up to the task?

EdSkyQuery-G: Motivation & Aims

- Support the “Open SkyQuery Initiative”
  - Move from a .NET-specific implementation.
  - Enable similar functionality on other platforms.
- Extensible framework for e-science:
  - Handle heterogeneous archives.
  - ‘Plug in’ algorithms e.g. Nearest Neighbour.
  - Interact with Astrogrid components & VO.
  - Leveraged for the BRIDGES project to perform simple joins.
- Apply Eldas to large scale E-Science problems:
  - Test: functionality, scalability & performance.
- Cross team collaboration:
  - Dr. Bob Mann (UoE, ROE), Edikt, EPCC, NeSC.

SkyQuery: High Level Architecture



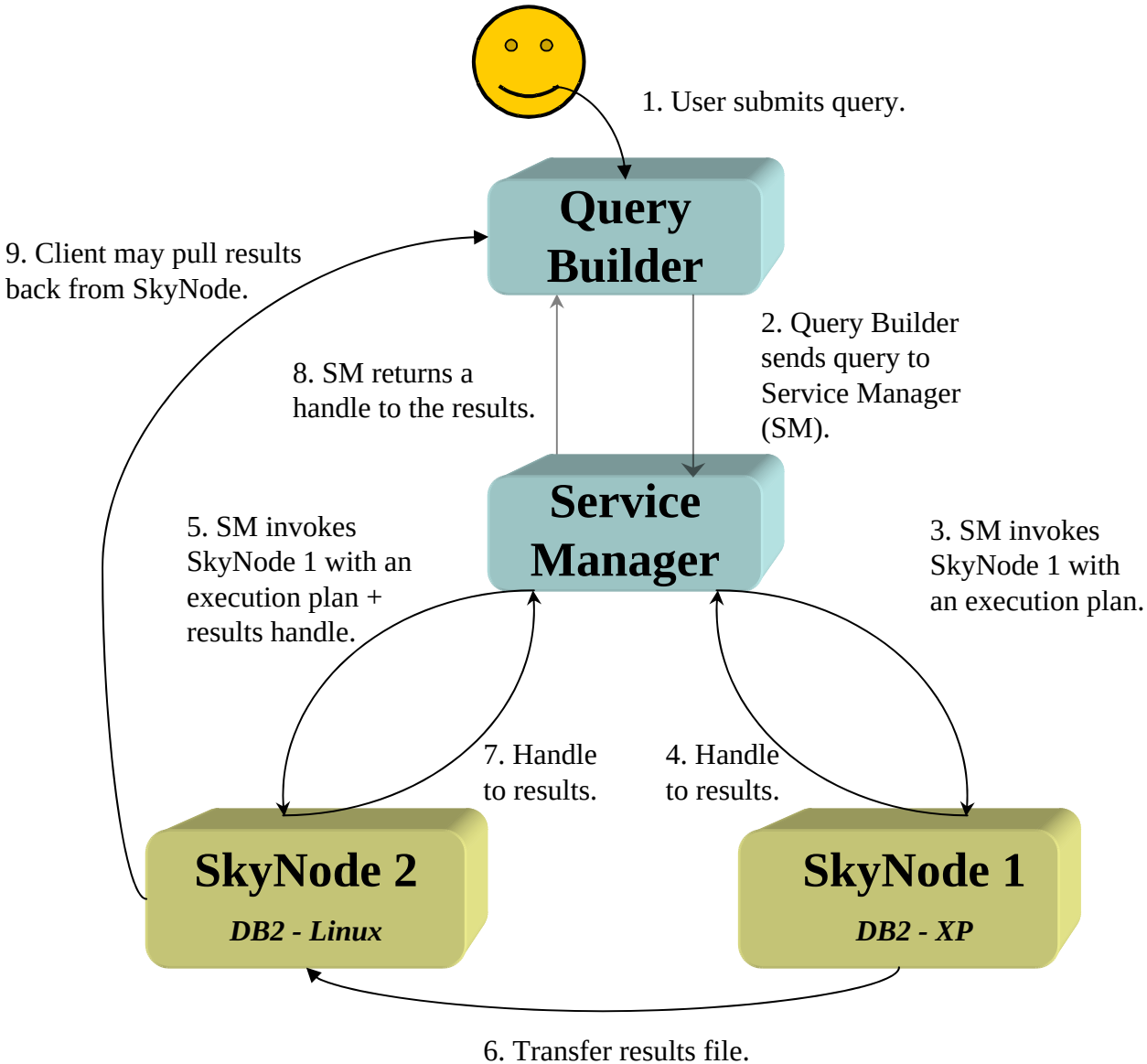
EdSkyQuery-G: Architecture

- Inspired by Greg Riccardi's paper for DAIS-WG:
  - http://www.cs.fsu.edu/~riccardi/grid/skyquery.pdf

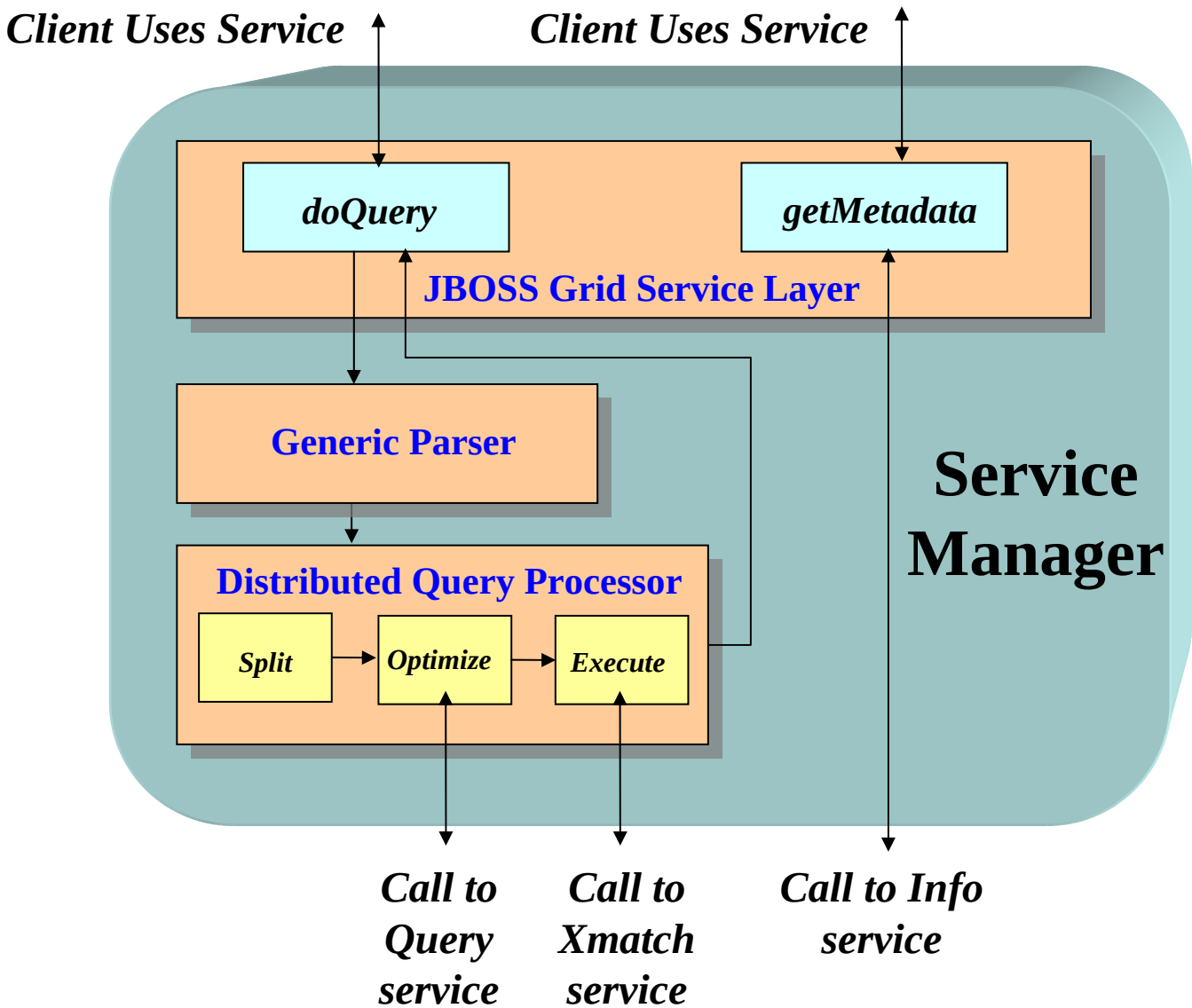
- Discusses two approaches:
  1. Access recipes for service interactions.
  2. Retained state for service interactions.

- Potential benefits of #2:
  - *Scalability*
  - *Robustness*
  - *Usability*

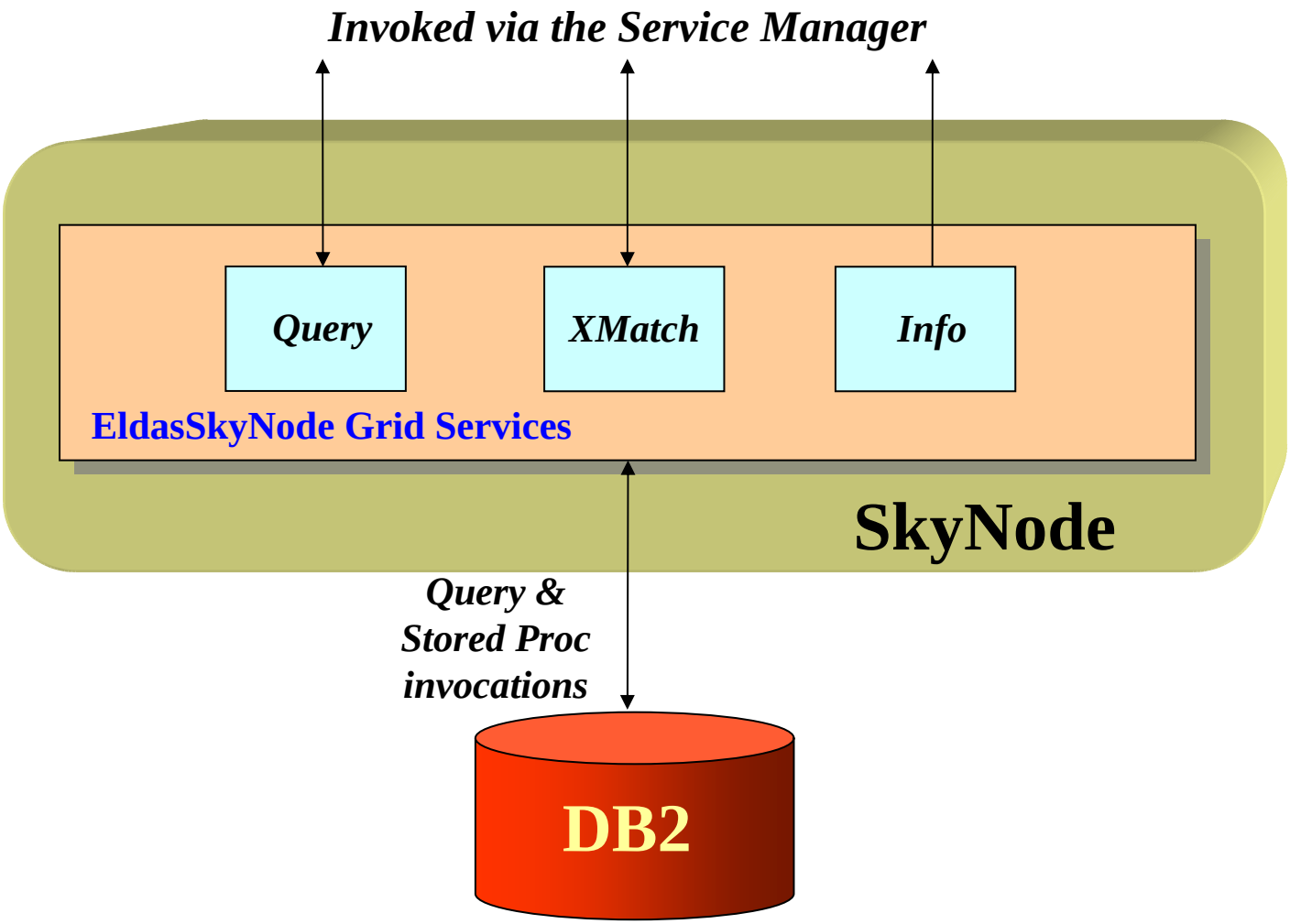
EdSkyQuery-G: Architecture



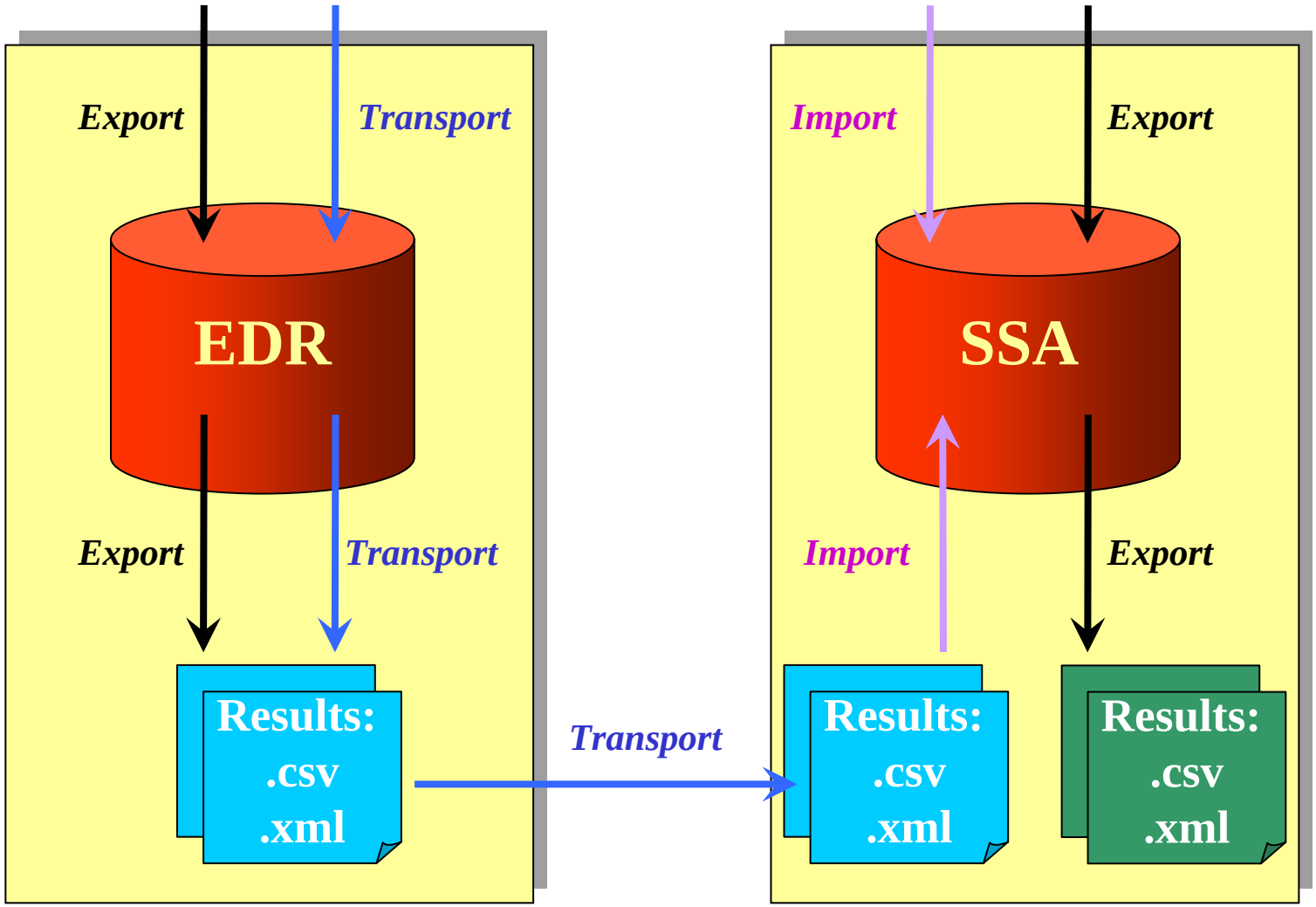
EdSkyQuery-G: Service Manager



EdSkyQuery-G: SkyNode Architecture



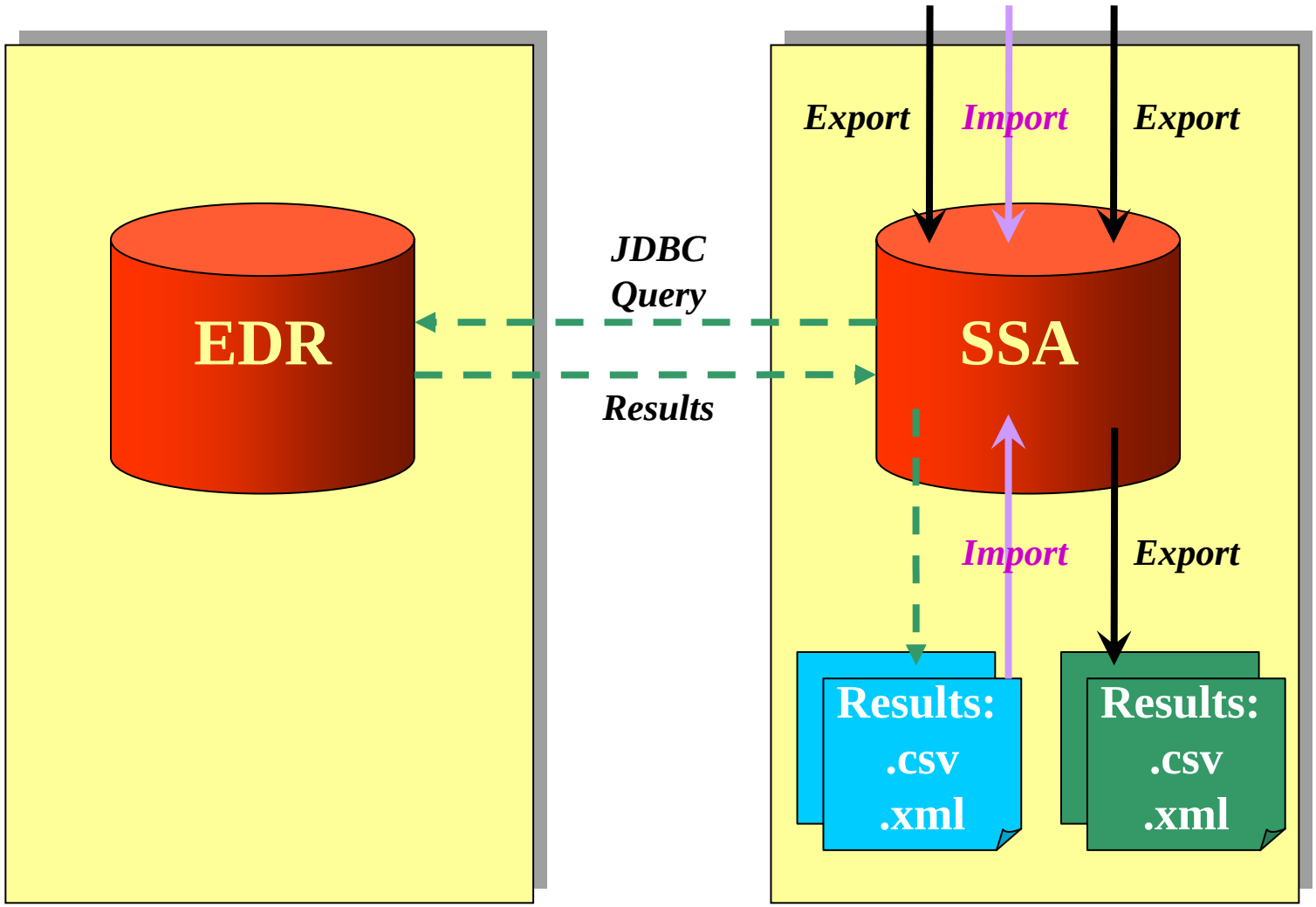
EdSkyQuery-G: Stored Procedures



Current Status

- Pre-Alpha (internal release), November ‘04.
- End-to-end invocation of all components:
  - Client->ServiceManager->SkyNode->Database
- 2 * 10GB DB2 test databases:
  - SSA from SuperCosmos, hosted by NeSC.
  - EDR from SDSS, hosted by EPCC.
- Limitations:
  - SM: No query parser/splitting.
  - Simple cross database join: not yet using XMatch.
  - No data transport, other than JDBC between databases.
  - Final results reside on database server.

EdSkyQuery-G: Pre-Alpha Stored Procedures



Results

- Tested with 3 queries from ROE cookbook:
  - http://surveys.roe.ac.uk/ssa/sqlcookbook.html
  - Queries #16, #17, #19.
- Results show:
  - Exporting data is quick.
  - Importing data is >10 * slower:
    - Should we use native DB calls rather than JDBC for import only?
  - Queries slow:
    - Need more indexes and database tuning?

Query No.	#Rows selected	Export Time (secs)	Import Time (secs)	Join Query Time (secs)	Total Time (secs)
16	488,718	130	1585	1110	2825
17	383,672	96	1138	1470	2704
19	4,667	82	15	1474	1571

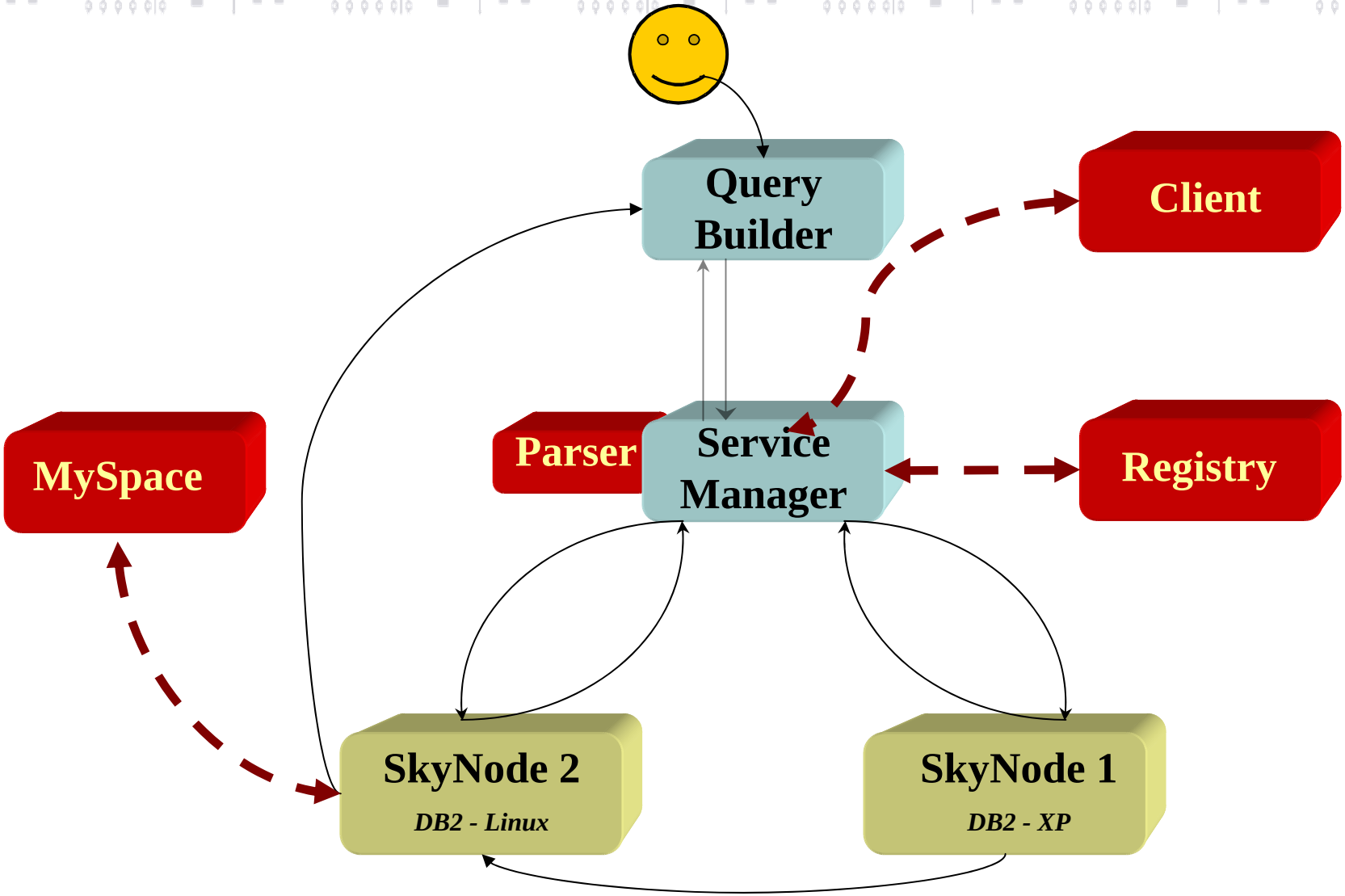
Deliverables

- Software (Internal):
  - Prototype client & GUI client.
  - Service Manager.
  - Eldas + Skynode Interface.
  - Database stored procedures (Java).
- Documentation (some on NescForge):
  - Use Cases, Requirements, Design.
  - Performance Testing, Installation.
- Papers
  - AHM 04 Poster & Paper:
    - http://www.allhands.org.uk/proceedings/papers/123.pdf
  - ADASS 04 Paper.

Short Term Focus

- Enable science:
  - Compare different cross match algorithms.
  - Incorporate XMatch, CSIRO, ROE as stored procedures.
- Data Transfer – SCP:
  - Between databases.
  - Pull results back to client.
  - Deliver to a third party.
- Client & Service Manager enhancements:
  - Handle broader range of queries.
- Performance:
  - Compare with pre-alpha benchmarks.
- Improve test infrastructure.

Longer Term Focus: AstroGrid



Longer Term Focus

- Interaction with Astrogrid components:
  - Clients, Registry, ADQL Parser, MySpace.
- OpenSkyQuery:
  - Compliance with interfaces.
  - Test with other DBMS and catalogues.
- Query Builder GUI/Data Integration tool:
  - Lead user through choosing fields from different datasets.
- GridFTP:
  - Currently unsuitable….
  - NeSC course in Jan 05 may reveal more.

edikt

National e-Science Centre

Thank you

Questions?