Databases for data management in PHENIX

Irina Sourikova

Brookhaven National Laboratory

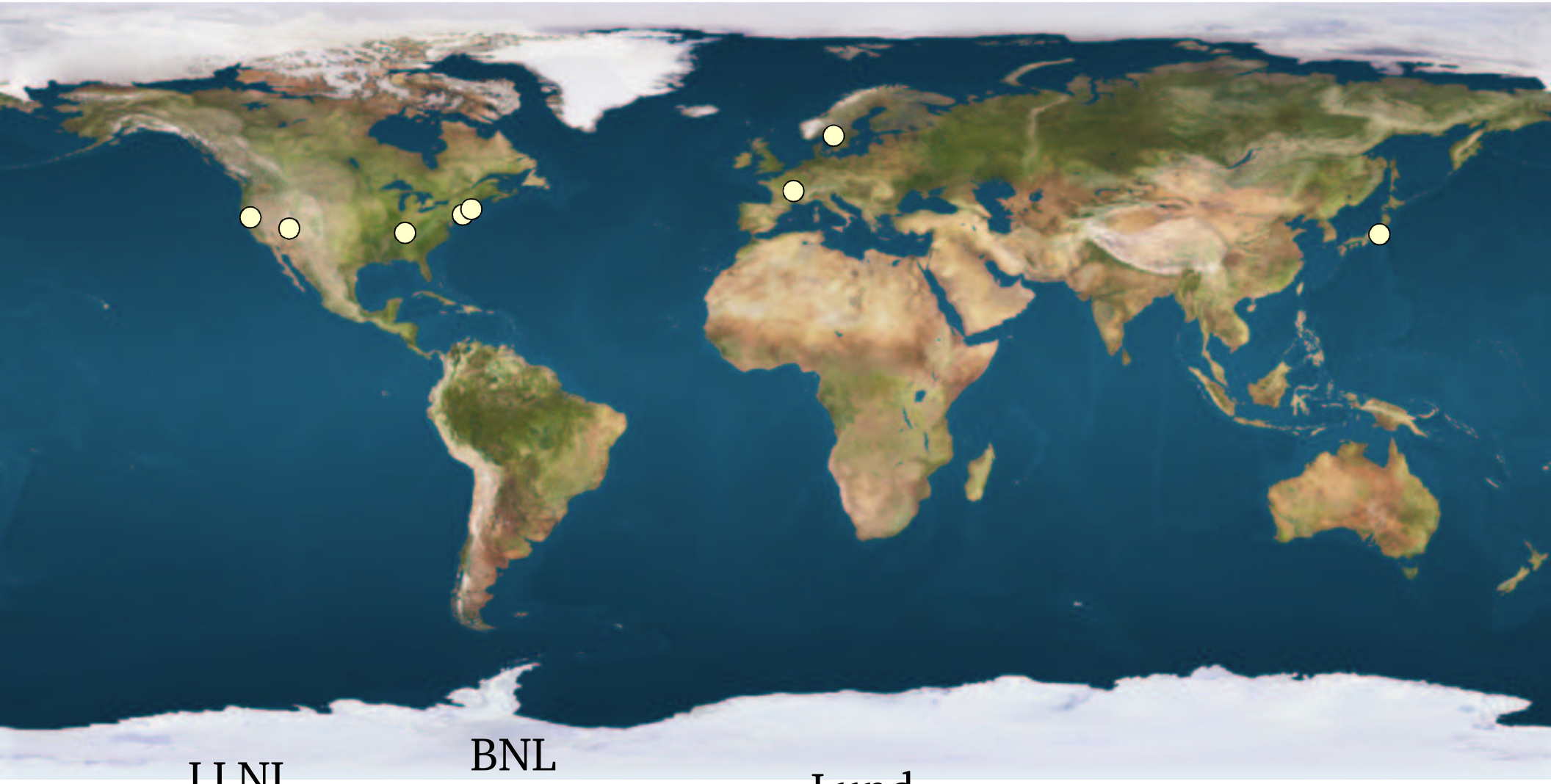
for the

PHEN  **ion**

PHENIX experiment

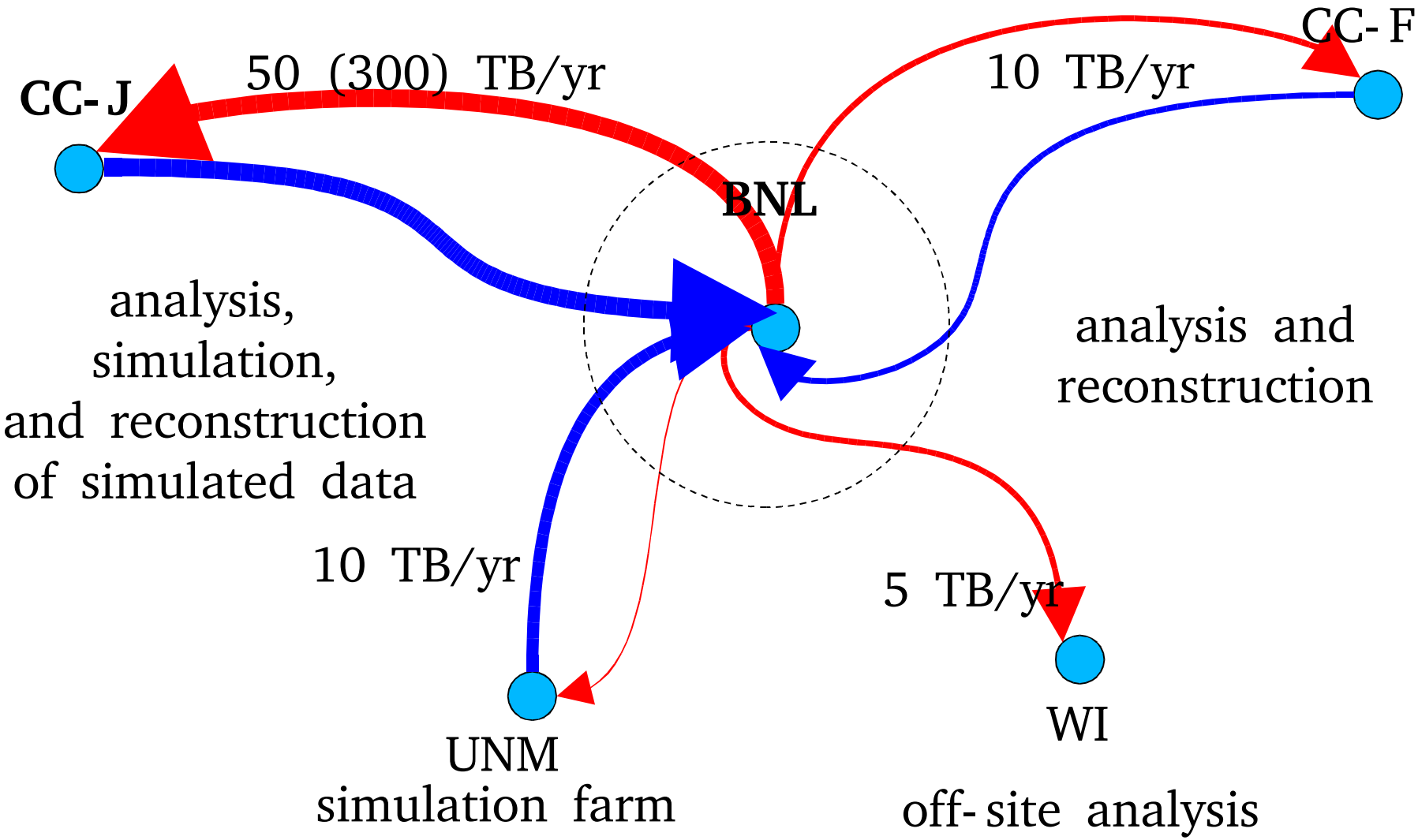
- PHENIX is a large RHIC experiment producing about 100TB of data a year
- 400+ collaborators in 57 institutions
- Widely distributed – 12 countries, 4 continents

Major PHENIX Computing Sites



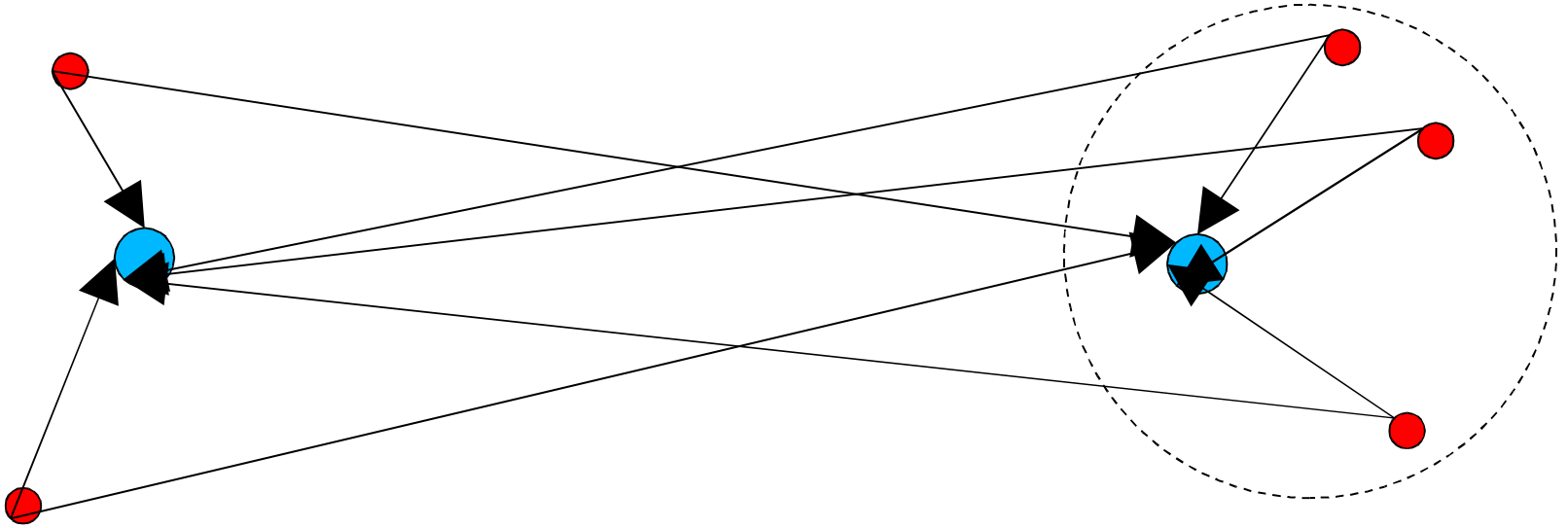
PHENIX data

- PHENIX is in the third year of data taking and has accumulated about 200,000 data files so far. At the end of current run (May 2003) this number will increase up to 400,000 files
- Files are produced at multiple computing centers
- Large amount of data is moved from/to BNL and between PHENIX institutions
- More on PHENIX data management in Andrey Shevel's talk in Category 2, 5:30p.m.

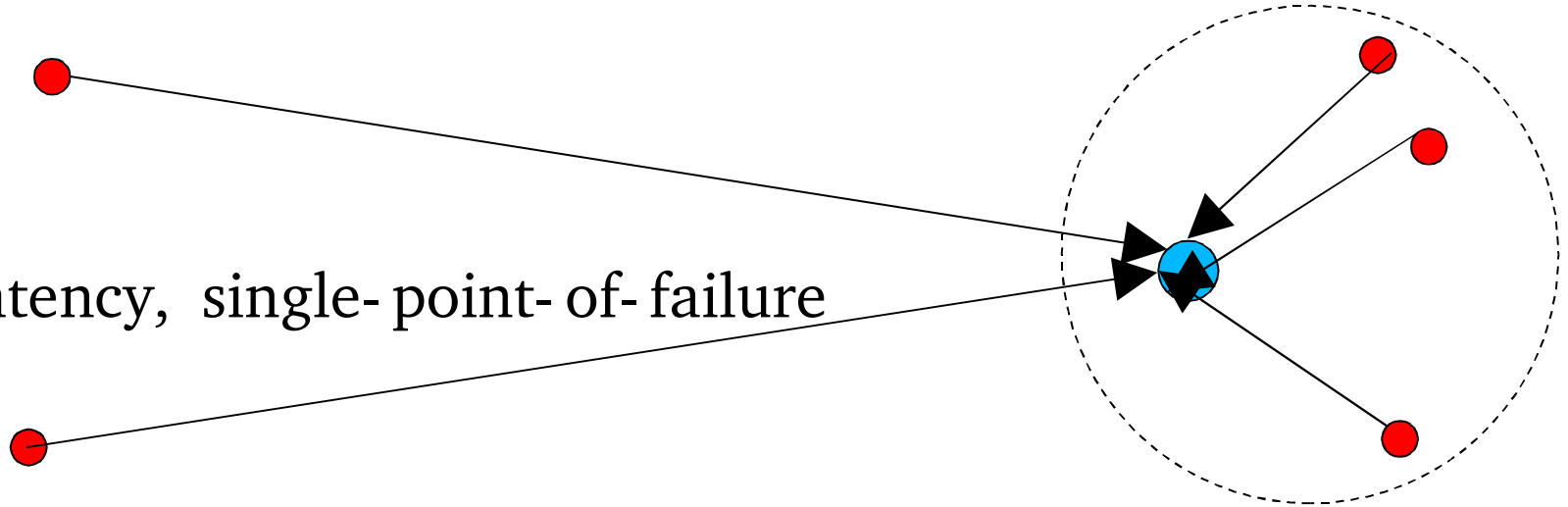

CC- J
50 (300) TB/yr
CC- F
10 TB/yr
BNL
analysis,
simulation,
and reconstruction
of simulated data
analysis and
reconstruction
10 TB/yr
5 TB/yr
WI
UNM
simulation farm
off- site analysis

File Catalog requirements

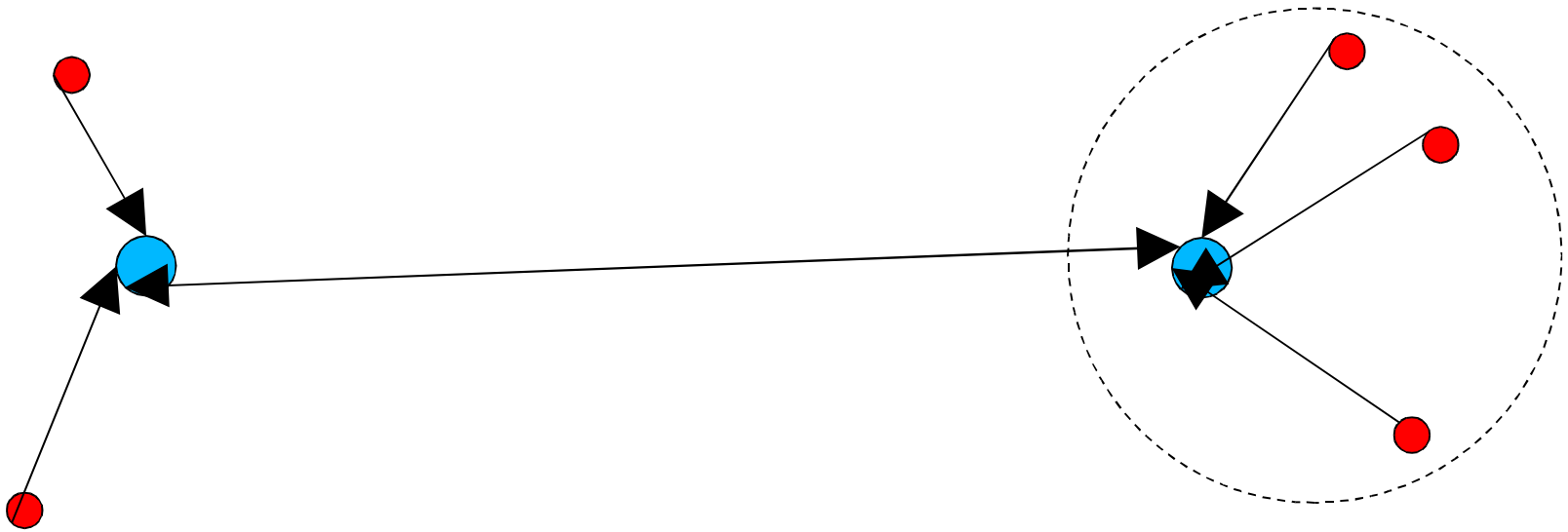
- Only 10% of the data can be kept on disk for analysis
- Set of disk resident files is constantly changing
- File Catalog should
  - Contain up-to-date data for both master and replica file location
  - Provide fast and reliable access at all PHENIX sites
  - Provide write permissions to the sites that store portions of the data set
- Last requirement was not satisfied by existing catalogs(Objectivity based PHENIX catalog, MAGDA, Globus Replica Catalog and others)



Objy: difficulty of PDN negotiation, exposes primary DB to world



WAN latency, single-point-of-failure



What we need: hybrid solution, extra software involved

Database technology choice

- Objectivity – problems with peer-to-peer replication
- Oracle was an obvious candidate(but expensive)
- MySQL didn’t have ACID properties and referential integrity a year ago when we were considering our options. Had only master-slave replication
- postgreSQL seemed a very attractive DBMS with several existing projects on peer-to-peer replication
- SOLUTION: to have central Objy based metadata catalog and distributed file replica catalog

PostgreSQL

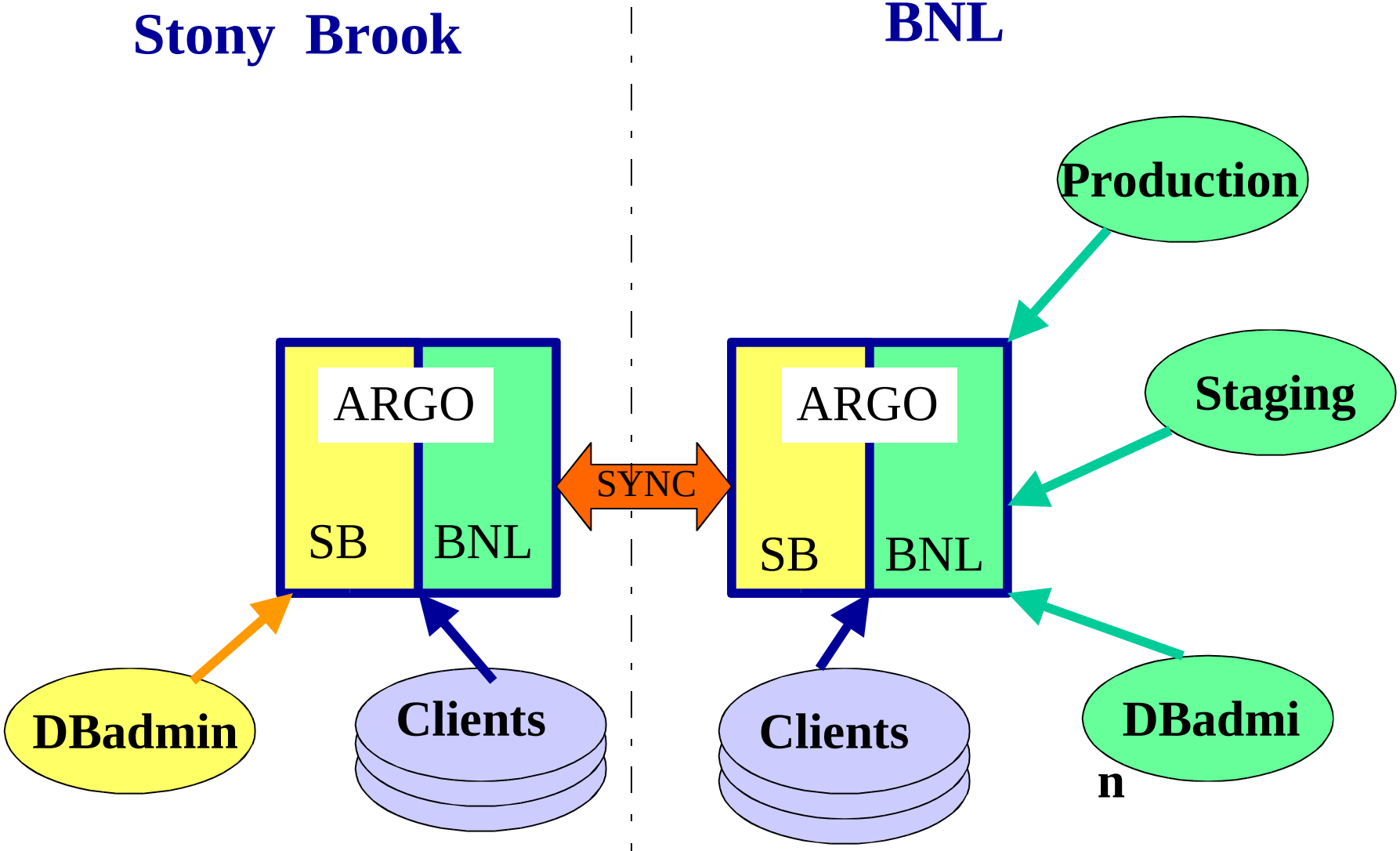
- ACID compliant
- Multi-Version Concurrency Control for concurrent applications
- Easy to use
- Free
- LISTEN and NOTIFY support message passing and client notification of an event in the database. Important for automating data replication

PostgreSQL Replicator

- http://pgreplicator.sourceforge.net
- Partial, peer-to-peer, async replication
- Table level data ownership model
- Table level replicated database set
- Master/Slave, Update Anywhere, Workload Partitioning data ownership models are supported
- Table level conflict resolution

Distributed PHENIX Catalog

- Distribution includes BNL and Stony Brook University
- Two sites have common tables that are synchronized during replica session
- Common tables have Workload Partitioning data ownership model, that means the table entries can be updated only by the site-originator


Stony Brook
BNL
Production
Staging
ARGO
ARGO
SYNC
SB
BNL
SB
BNL
DBadmin
Clients
Clients
DBadmi
n

File Catalog replication

- Synchronization is two-way
- DB replication is partial, only latest updates are transferred over the network
- That solves scalability problems
- Replication of 20 000 new updates takes 1min

What we gained by distributing File Catalog

- Reliability(no single point of failure)
- Accessibility(reduced workload, no WAN latency)
- Solved security problems (firewall conduits only between servers)
- Catalog data is more up-to-date(all sites update the catalog online)
- Less manpower needed to maintain the catalog