Global IT Innovator
NTT DATA GROUP

PostgreSQL and Postgres-XC in NTT Group

Nov. 3rd, 2011
Koichi Suzuki
NTT DATA Intellilink

Agenda

NTT DATA INTELLILINK CORPORATION

- Introduction to NTT group
  - Business
  - Major group companies
- PostgreSQL and NTT group
  - Why PostgreSQL?
  - Support and development organization
  - Contribution to PostgreSQL
  - PostgreSQL use --- present and future
- Postgres-XC and NTT group
  - Why large scale cluster?
  - How XC scales
  - How XC works
  - Current status and near future plan
  - New members wanted

NTT DATA INTELLILINK CORPORATION

Introduction to NTT group

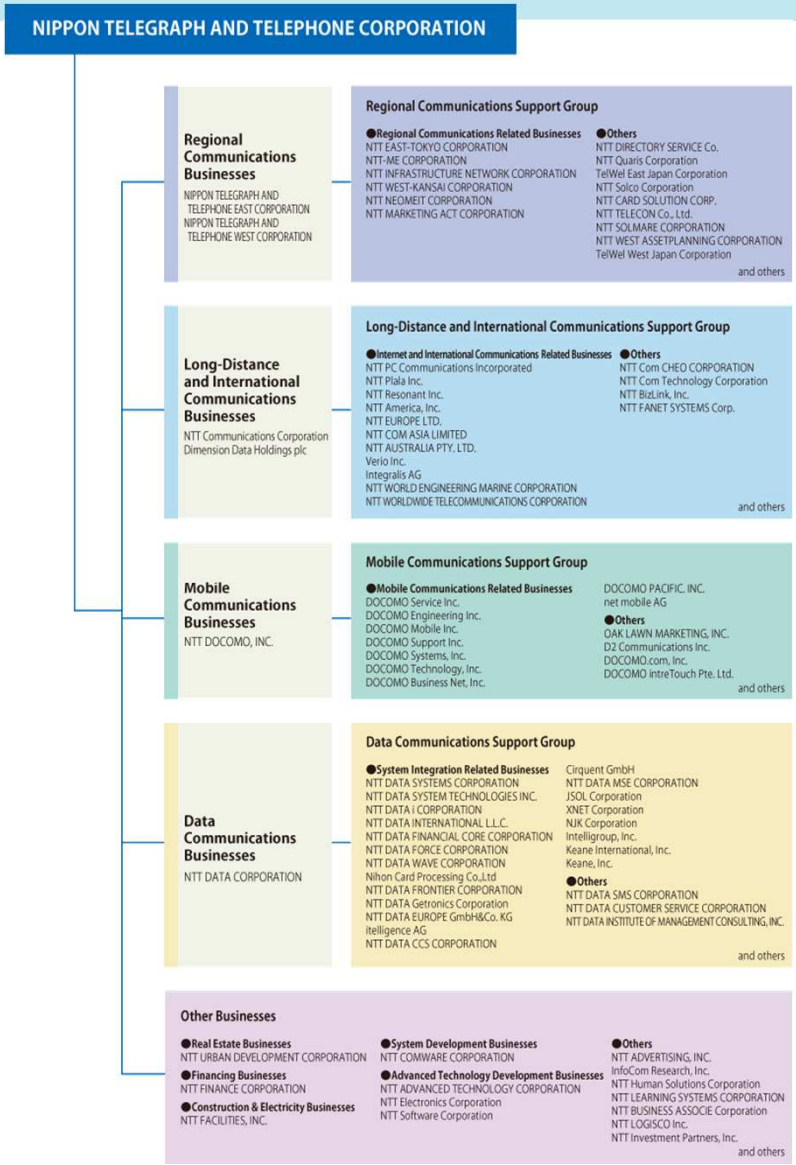
NTT Group

NTT DATA INTELLILINK CORPORATION

- Second Largest Telecommunication Company
  - Local Service
  - Long Distance/Overseas Service
  - Mobile Service
  - System Integration
  - R&D Facility
    - Holding Company
    - Each Member Company

NTT Group (cont.)

NTT DATA INTELLILINK CORPORATION



http://www.ntt.co.jp/about_e/group2.html
http://www.ntt.co.jp/about_e/group.html

Corporate Data

NTT DATA INTELLILINK CORPORATION

Holding Company

Name	NIPPON TELEGRAPH AND TELEPHONE CORPORATION
Date of Establishment	April 1, 1985 In accordance with the Nippon Telegraph and Telephone Corporation Law （Bill No. 85, December 25, 1984）
Number of Employees	2,900（As of March 31, 2011）

As A Group

Total Assets:	¥19.6656 trillion ($260 billion)
Number of Employees:	219,350
Operating Revenues:	¥10.1814 trillion ($133 billion)
Number of Consolidated Subsidiaries:	756

1$ = ¥76.84

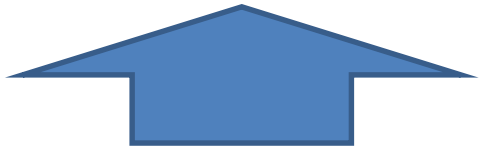
NTT DATA INTELLILINK CORPORATION

PostgreSQL and NTT Group

NTT Group and Open Source

NTT DATA INTELLILINK CORPORATION

- Total Cost of Ownership
- Longer support period
- Quick problem fix
- Expedite Open Source Software Deployment



Open Source Software Center (OSSC)

- **Collected more than one hundred open source engineers**
- **Established dedicated organization in April, 2006**

NTT OSSC Coverage and Acitvities

NTT DATA INTELLILINK CORPORATION

Coverage

- From kernel to integration
  - Linux Kernel
    - Distribution Support
  - Web Server
  - JBoss
  - PostgreSQL
  - Hadoop
  - Recommended Integration

Activities

- Support
- Consultation
- Evaluation
- Provide technical information
- Development through Communities

About Myself

http://www.intellilink.co.jp/plan/corporate/fellow_OSS-DB.html

NTT DATA INTELLILINK CORPORATION

- Belong to NTT DATA Intellilink
- Working for NTT OSSC
- Leader and Architect of Postgres-XC






NTT OSSC Location

NTT DATA INTELLILINK CORPORATION

150 Yrs. Ago

Present

Shinagawa Area: one of the transportation hub in Tokyo

NTT DATA INTELLILINK CORPORATION

PostgreSQL Deployment

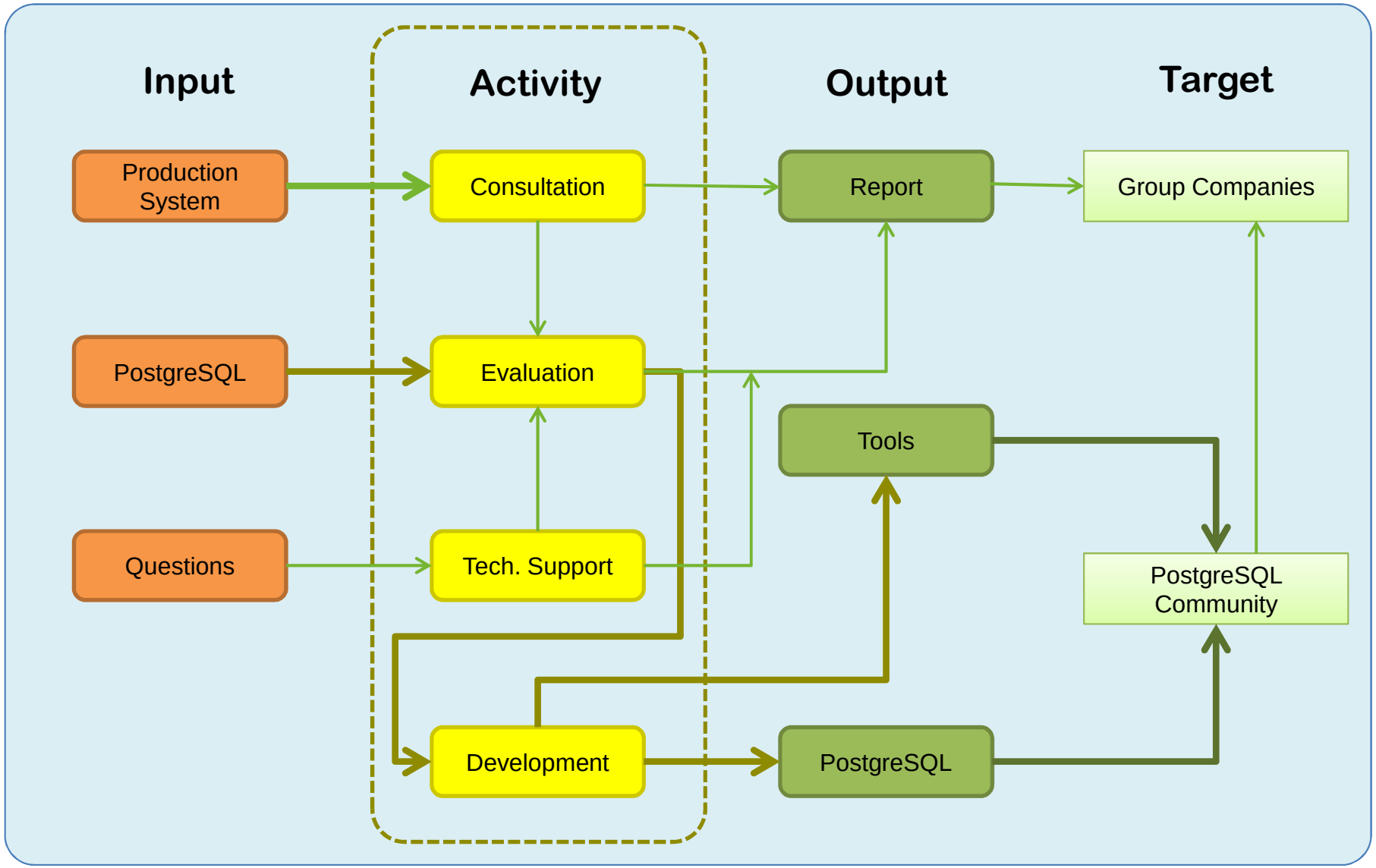
Understanding User Needs

NTT DATA INTELLILINK CORPORATION

- Information on performance
  - Show good and stable performance
  - Availability/reliability
    - downtime to recovery (e.g. 5min/yr for five-9s)
  - To prepare equipment (HDDs, CPUs etc.)
- Operation capability
  - compatibility with other operation tools
  - Usability
- Improve performance and usability
- Technical support

PostgreSQL Activities in OSSC

NTT DATA INTELLILINK CORPORATION



Evaluations

NTT DATA INTELLILINK CORPORATION

- What characters are important?
  - Most systems are OLTP not OLAP
  - Types of Transactions; read/write intensive
- TPC C and TPC W models are used
  - C model (DBT-2): write, I/O intensive
  - W model (DBT-1): read, CPU intensive
  - Other models: pgbench, DBT-3
- Throughput and stability
  - Peak performance test (3Hr. Workload > 90%)
    - CPU scalability
  - Long-run test (72Hr. 70% workload)
    - observe stability during vacuum and checkpoint

Throughput Result

NTT DATA INTELLILINK CORPORATION

- **Results of PostgreSQL and other DBMS.**
  - **Helped to adapt PostgreSQL for production systems with particular population and frequent requests.**

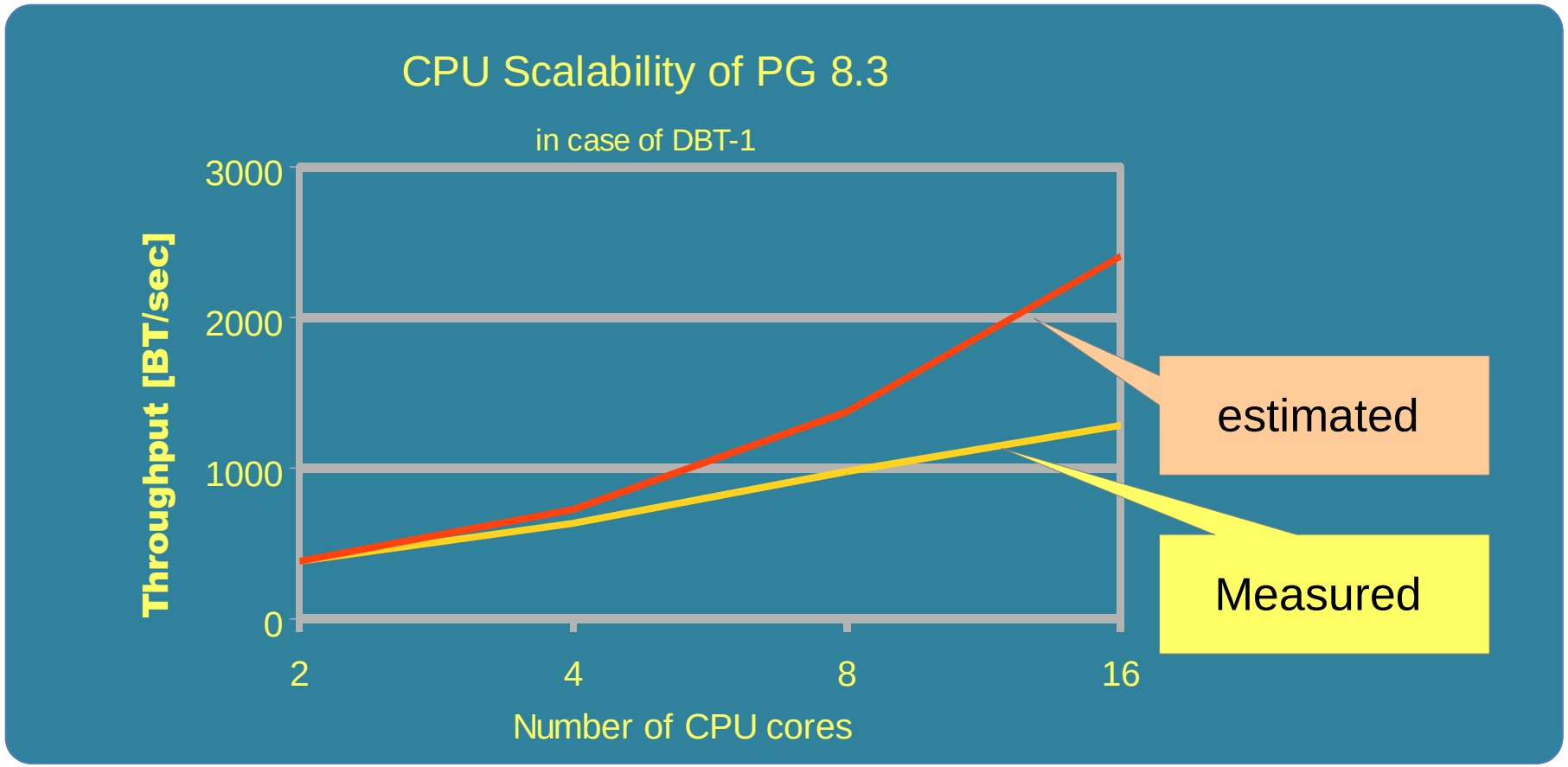
	8.2	8.3
TPC-W WIPS rd:wrt = 8:2	155% (1700tps)	190% (2100tps)
TPC-W IPSo rd:wrt = 5:5	80% (1100tps)	150% (2100tps)
TPC-C rd:wrt = 1:9	45% (123tps)	60% (165tps)

Equiments used for evaluations;
[TPC-W] Server: HP DL380G5 (Xeon 5160 3GHe, 12GB memory), Storage HP MSA500
[TPC-C] Server: DL580G4(Xeon DC 3.4 GHz 4 core, 24GB memory), Storage HP MSA 1000
[OS] Redhat Enterprise Linux 5 update 1
Values are gotten from 48 hours execution and displayed in average.

CPU Scalability

NTT DATA INTELLILINK CORPORATION

- CPU and servers tend to have more cores
  - 4-8 for middle-scale, 32 for large-scale.
  - Good scalability up to 8 cores for 8.3 and later.



Throughput Evaluation Result

NTT DATA INTELLILINK CORPORATION

- Relatively good performance compared with other DBMS.
  - Helped choosing PostgreSQL for production systems having particular population and frequent requests.
  - PostgreSQL is feasible to replace proprietary DB
- Average performance is sufficient
  - How about in extreme case? How is it stable?
    - Stability of performance

Perfomance Stability

NTT DATA INTELLILINK CORPORATION

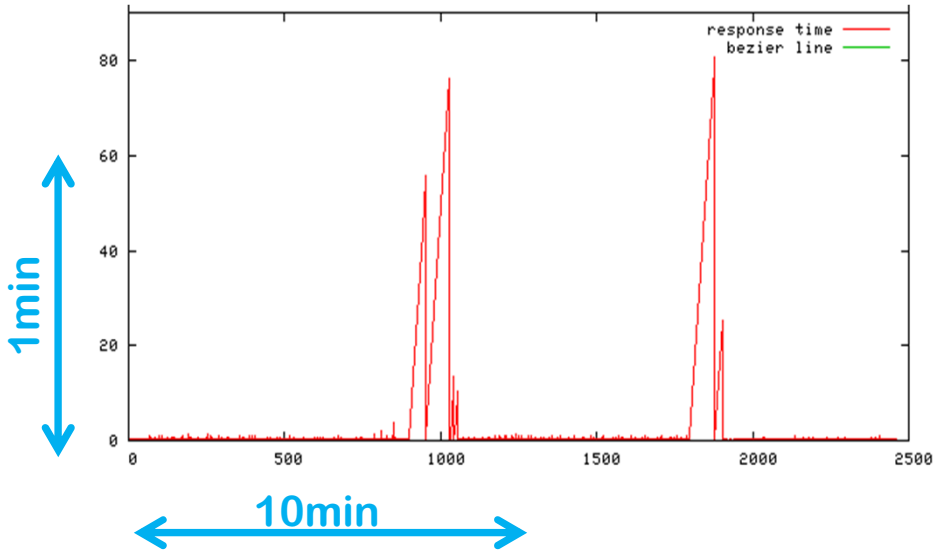
- Performance stability is important
  - Avoid queries keep executing for a long time
  - Can guarantee minimum performance (e.g. longest response time)?
- Observe stability with long-run test.
  - Vacuums and checkpoints done many times
  - Long-run stability evaluated with TPC-W
    - Workload itself stable against time
    - TPC-C increases data population and (in result) workload while running.

Stability Test (1)

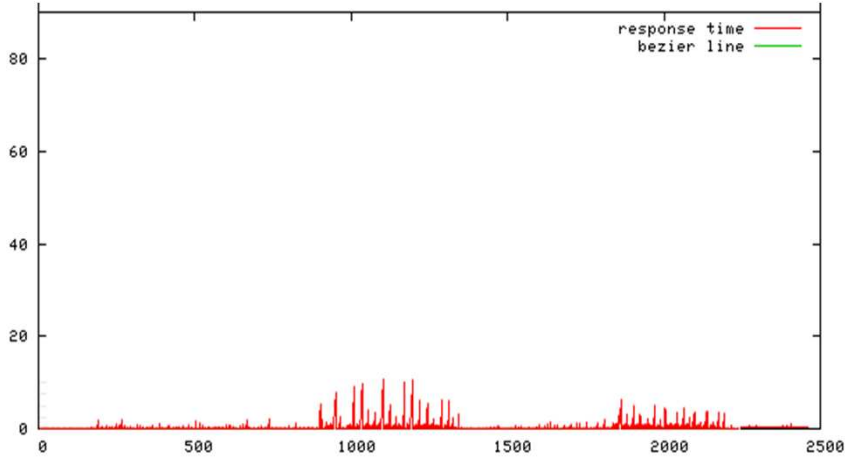
NTT DATA INTELLILINK CORPORATION

- Found PostgreSQL 8.3 performance is significantly stable compared with 8.2
  - PostgreSQL 8.2 (Left) glitches caused by checkpoints
  - PostgreSQL 8.3 (Right) glitches reduced 20% of 8.2
  - Glitches in 8.2 concerned to be obstacle for production systems.

PostgreSQL 8.2



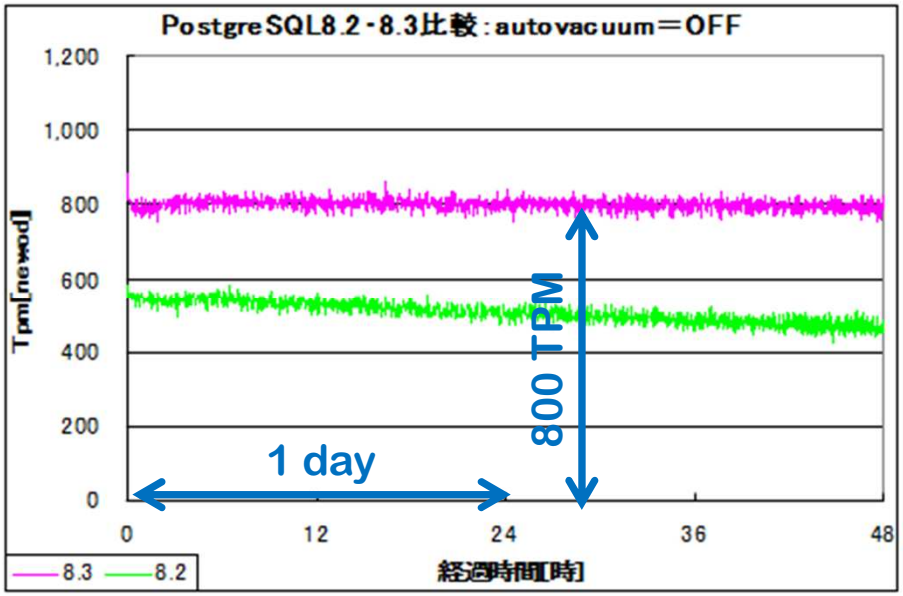
PostgreSQL 8.3

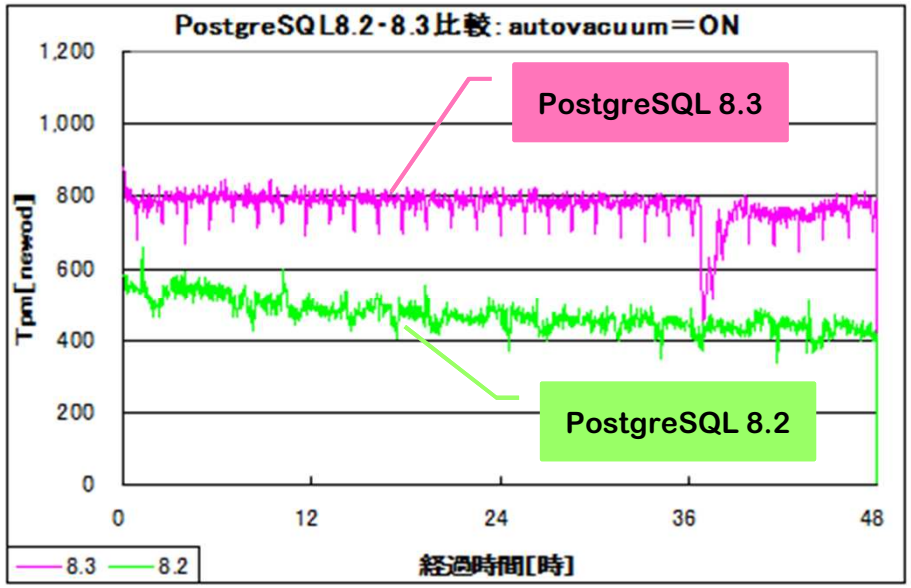


Stability Test (2)

NTT DATA INTELLILINK CORPORATION

- Influence of dead tuples and vacuum op.
  - autovauum=off (Left) in PostgreSQL 8.2 reduces performance
  - autovauum=on(Right) both caused glitches



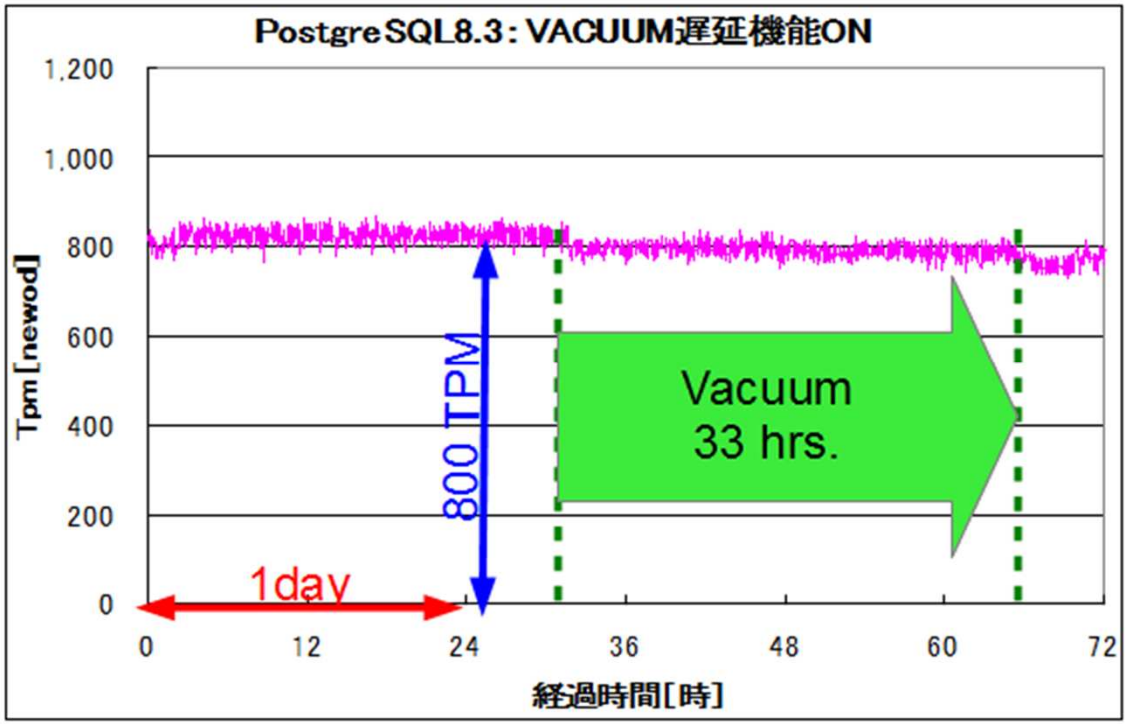


http://lets.postgresql.jp/documents/case/ntt_comware/2

Stability Test (3)

NTT DATA INTELLILINK CORPORATION

- Cost-based vacuum works well
  - Cost-based vacuum smooths through put
    - Vacuum prolonged to 33 hrs from 2 hrs prev. case



http://lets.postgresql.jp/documents/case/ntt_comware/2

Evaluation Summary

NTT DATA INTELLILINK CORPORATION

- PostgreSQL 8.3 shows sufficient performance for NTT Group production systems with middle scale DB.
  - SInce version 8.3, deployment has been accelerated.
  - Vacuum with HOT and cost-based, time-spread checkpoint are important improvements.
    - Improved vacuum reduces operation design.
- Remaining issues...(including other evaluations)
  - More CPU scalability (e.g. 64 cores)
  - More efficient I/O handling (I/O bandwidth evaluation shows that PostgreSQL writes 4 times more than commercial DBMS)
  - Shorter recovery time.

Database Operation Evaluation

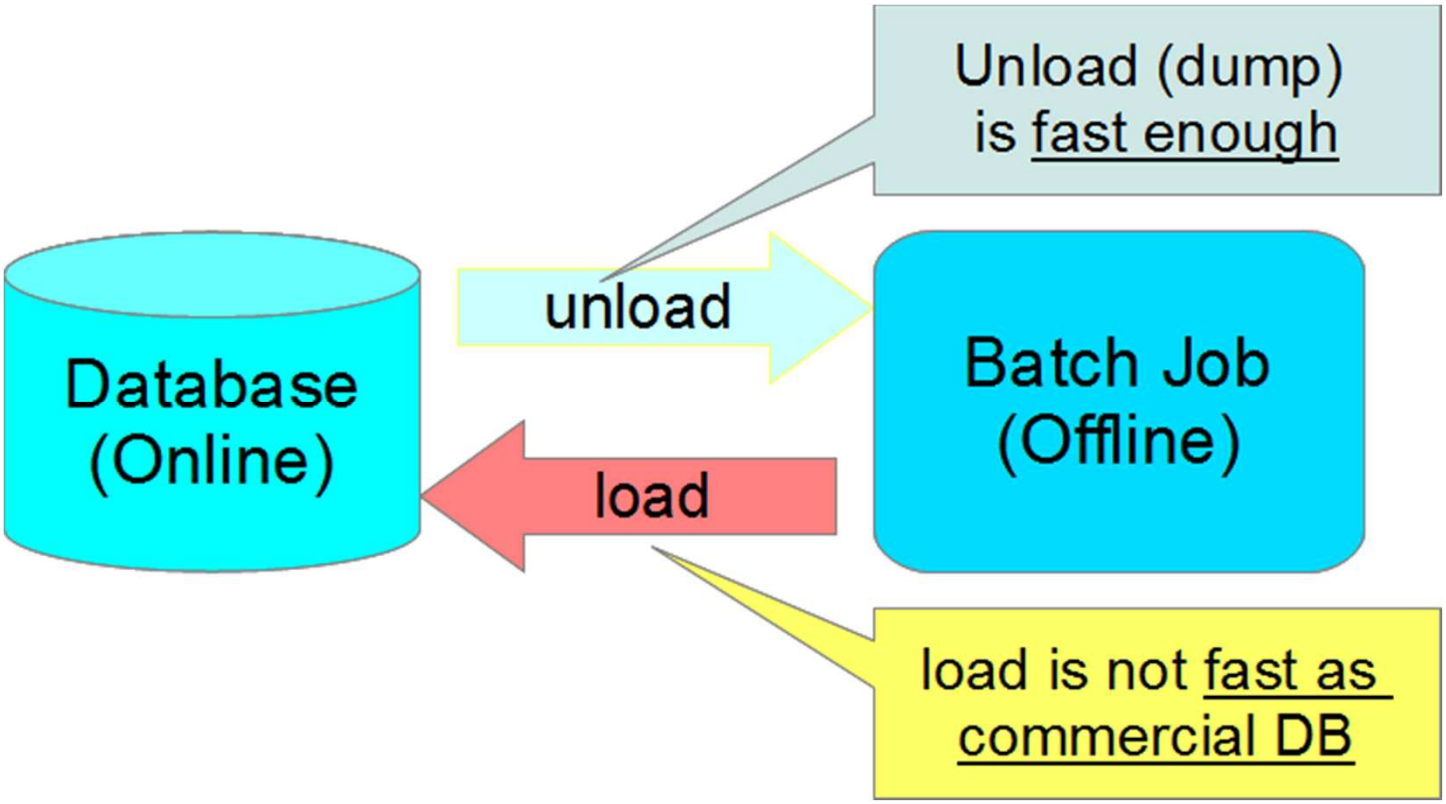
NTT DATA INTELLILINK CORPORATION

- Backups:
  - Logical: pg_dump is easy to use but not used widely in online operation (in NTT Group) because it is hard to tell what transactions are included in the dump.
  - Physical: PITR is nice, but operation is complicated and is not easy to use.
    - Need dedicated consultation or out-of-the-box package.
- Data loading:
  - COPY is useful but not fast enough.
  - Offline data loading can speed up daily batch jobs.

Database Operation Evaluation

NTT DATA INTELLILINK CORPORATION

- **Use of Fast Data loading:**
  - **DB migration for production system must be done in limited time period.**
  - **Offline data loading can speed up batch jobs (below).**



Monitoring

NTT DATA INTELLILINK CORPORATION

- Querying PostgreSQL internal statistics provides useful data for tuning and trouble shoot.
  - we need external tool that get and collect PostgreSQL's internal statistic data proactively.
  - Some troubles are difficult to reproduce. Queried data can be used for post-mortem analysis.

Target	Purpose	Means	Available?
Live or Dead	Fail over the server	Monitor process ID	Yes
Slow Query	Trouble shooting	Operation logs	Yes
Internal Statistics	Trouble shooting	Query to PostgreSQL	No dedicated tools

Development

NTT DATA INTELLILINK CORPORATION

- PostreSQL core
  - Stability
  - Availability
- Peripheral tools
  - Backup
  - Data loading
  - Monitoring tool

Performance stability

NTT DATA INTELLILINK CORPORATION

- NTT OSS Center donated some functionality for Vacuum and Checkpoints
  - Most of them were accepted to PostgreSQL core
    - Cost-based vacuum
    - multiple concurrent autovacuum processes
    - Checkpoints spread out (smooth checkpoint)
  - These help PostgreSQL performance stability, which accelerate introduction.

Availability Improvement

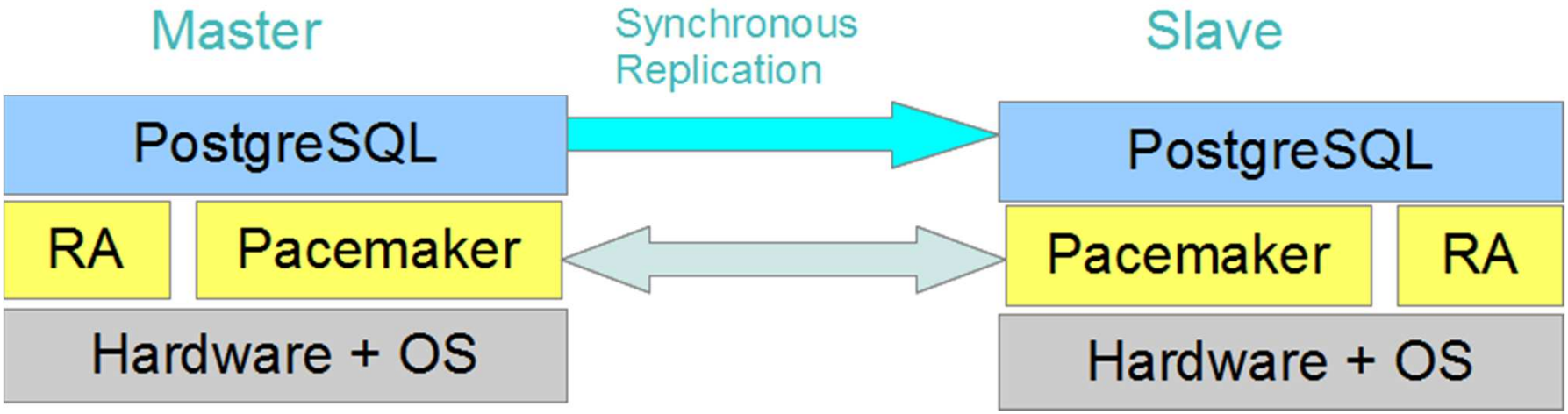
NTT DATA INTELLILINK CORPORATION

- About 1/3 NTT systems require fail over within 1 min.
  - Fail over cluster with shared disk requires fsck when swiching, which takes several minutes.
  - Replication clusters using query replication guarantee loss-less fail over, however impose incompatibilities with original PostgreSQL.
- We start to develop stream replication in 2006.
  - At first, proprietary product, then made OSS in 2008.
  - Proposal at PGCon 2008 (Mr. Fujii)
  - Streaming replication (asynchronous mode) was committed to 9.0 (2010)
  - Synchronous mode is now in 9.1

Availability Improvement (2)

NTT DATA INTELLILINK CORPORATION

- Peripheral software for HA has been developed
  - To switch server when failure, Linux-HA (Pacemaker) is used
    - We also uses Pacemaker for High-availability system
  - Pacemaker's Resource Agents for operation



HA Cluster Applications

NTT DATA INTELLILINK CORPORATION

- HA Cluster including PostgreSQL equipped with synchronous Replication is expected to be suitable for applications required more higher reliability;
  - Telecommunication support systems
  - Trading systems
  - Web commerce with high-availability

pg_rman ; backup tool

NTT DATA INTELLILINK CORPORATION

- Motivation ; FAQ.
  - PITR is powerful but complicated
    - When can we discard old archive logs?
    - How can we identify what archive logs are needed?
- Solution
  - Tool to automatize PITR operation
- Pg_rman
  - Collects all the files to needed to recover.
  - Works with one command.
  - Back-up files are organized into catalog.

http://code.google.com/p/pg-rman/

pg_bulkload; fast data loader

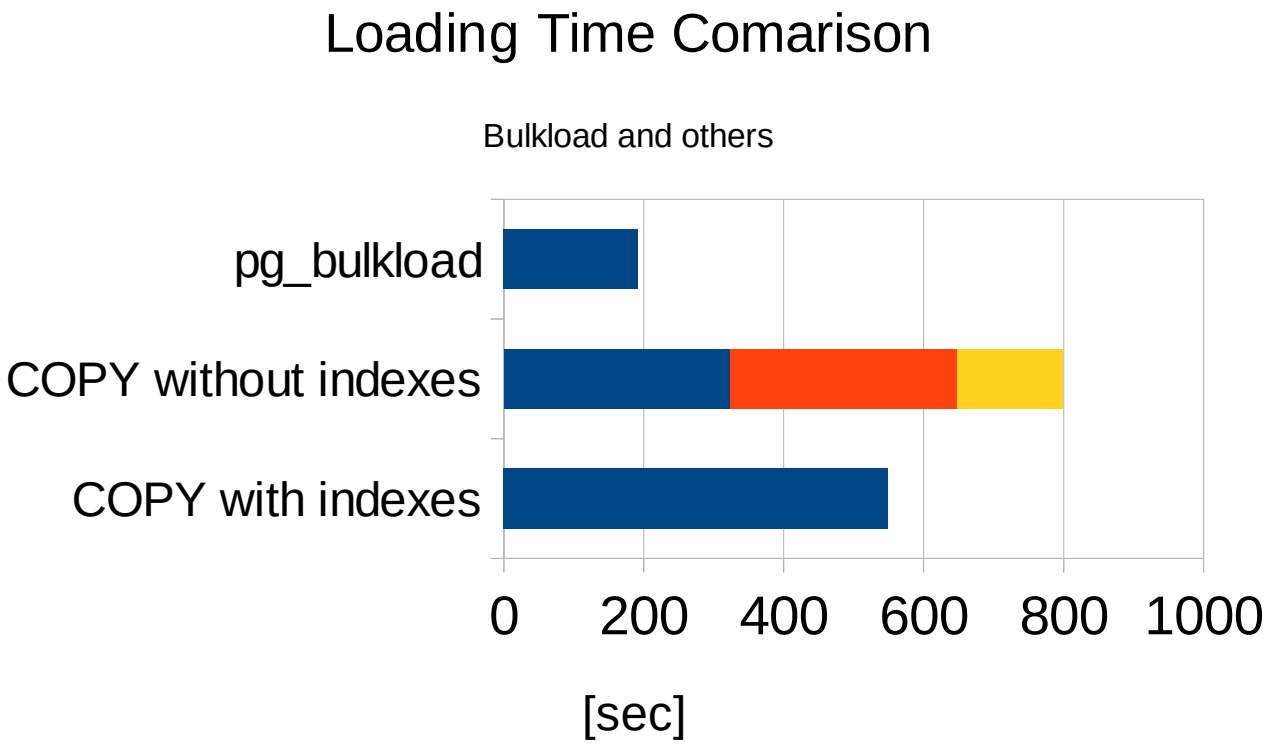
NTT DATA INTELLILINK CORPORATION

- Motivation: Data migration speed up.
  - Data migration in production systems should complete within scheduled time period
    - Data migration duration dominates DB size limit for PostgreSQL
    - COPY was not quick enough (ca. 2005)
- Solution
  - Dedicated Loading Tool; pg_bulkload
    - Initial and append modes
    - Direct and parallel load
    - Fast index creation

pg_bulkload; data loader(2)

NTT DATA INTELLILINK CORPORATION

- **Pg bulkload is as 2-3 times fast as COPY**



http://pgbulkload.projects.postgresql.org/index.html

pg_statsinfo; monitoring Tool

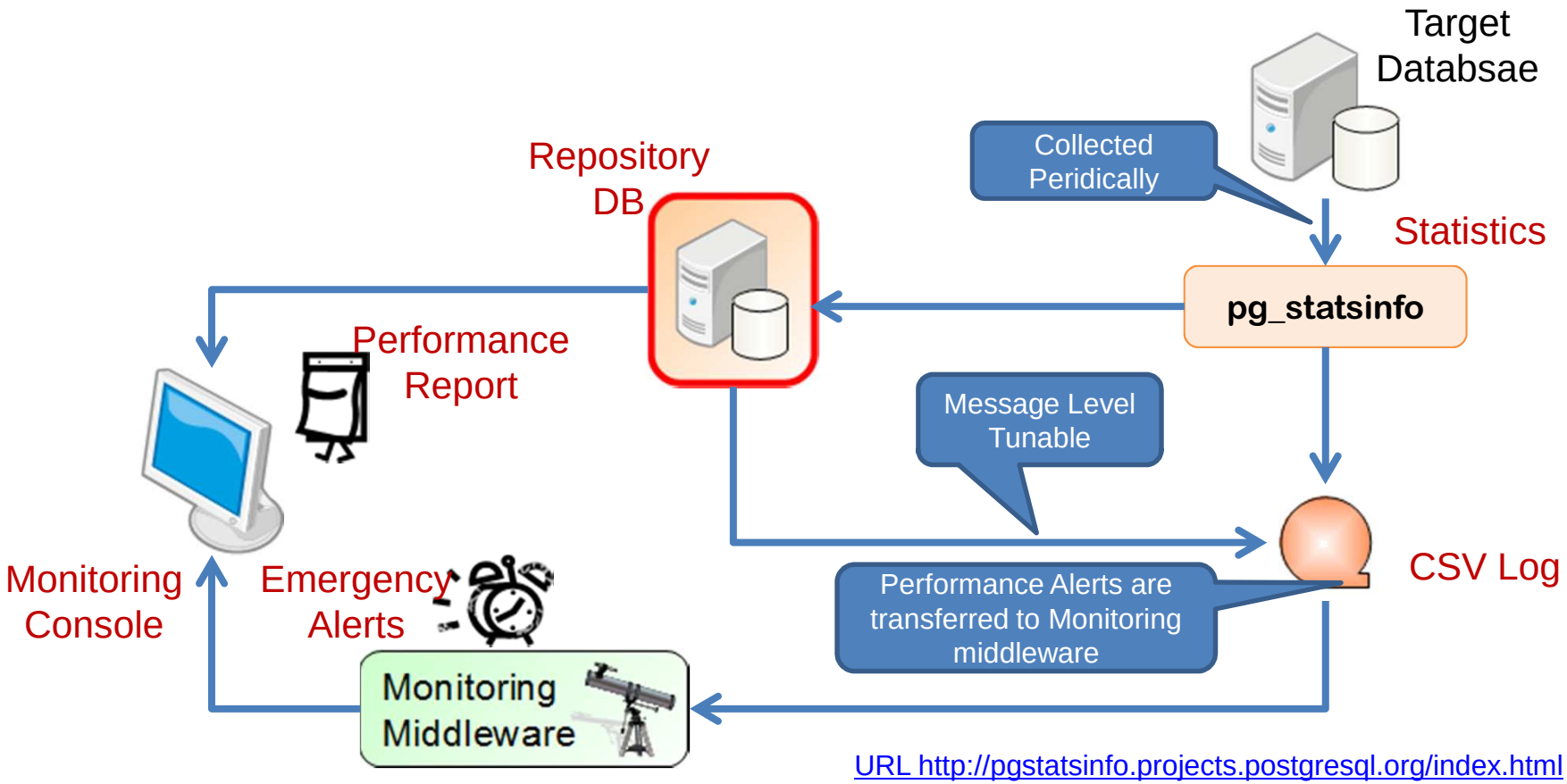
NTT DATA INTELLILINK CORPORATION

- Motivation
  - Effective support activity
    - Post-mortem analysis
  - Handy performance monitor
    - Predict performance trouble beforehand
- Features
  - Statistics collector with low power-consumption
    - Monitoring system runs (partially) on the Production system.
  - Visualize statistics
  - Programmable alert

pg_statsinfo; outline

NTT DATA INTELLILINK CORPORATION

- Collected data generate 'Report' and 'Alert'
  - Configuration: statistics collector + message filter for alert
  - Lower resource consumption: overhead < 3%



URL http://pgstatsinfo.projects.postgresql.org/index.html

Support Activities

NTT DATA INTELLILINK CORPORATION

- Technical Q and A
  - A few hundreds questions answered a year within 3 business days
  - Various questions
    - From usages to trouble issues
- Consultation
  - Migrate from Proprietary DBMS
    - Migration know-hows are cataloged (ca. 50 items; e.g. “how to rewrite synonym in Oracle”)
  - Performance tuning aids
    - Evaluate particular workloads and suggest tuning.

NTT Cases

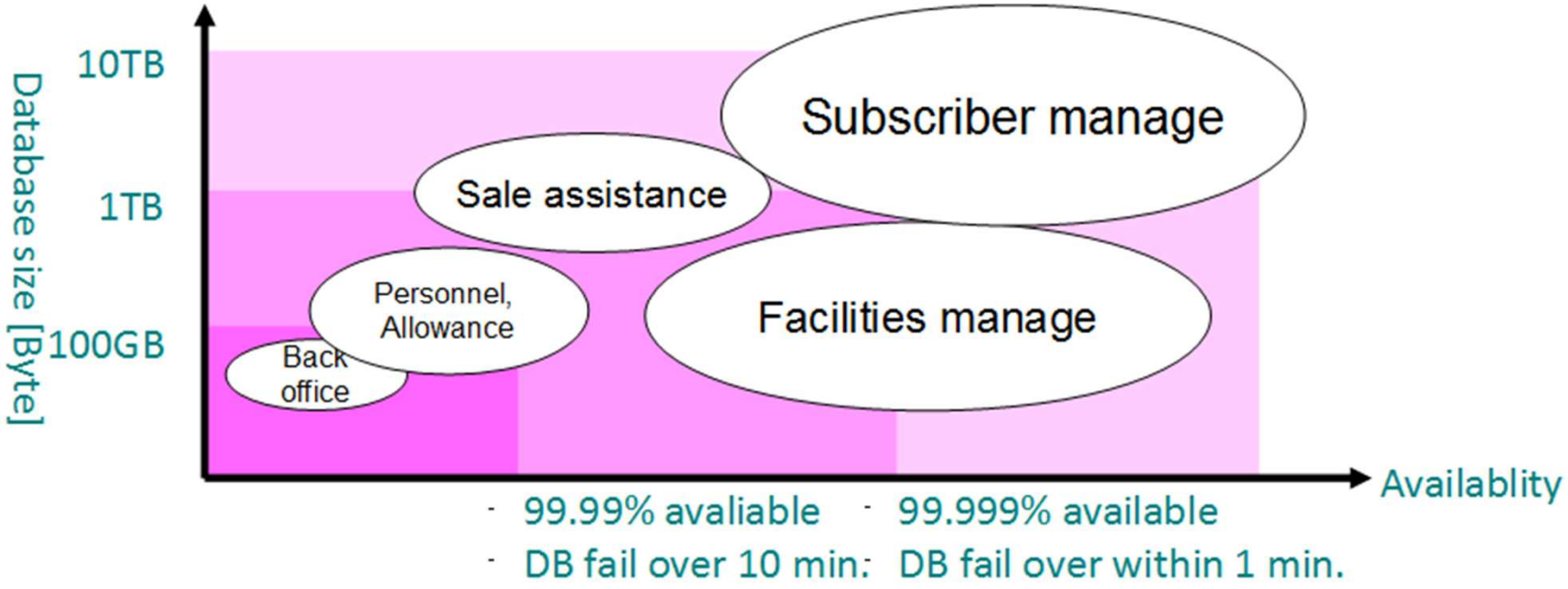
NTT DATA INTELLILINK CORPORATION

- OSS Center has introduced PostgreSQL to more than 100 systems; Highlight specs as follows
  - DB Size: Largest 3TB.
  - Frequency: 1000 TPS (or more)
  - HA: fail over takes less than 1 min. (15 sec. measured)
- Statistical Facts expressed
  - Individual cases are not allowed to disclose.

View of NTT's Production systems

NTT DATA INTELLILINK CORPORATION

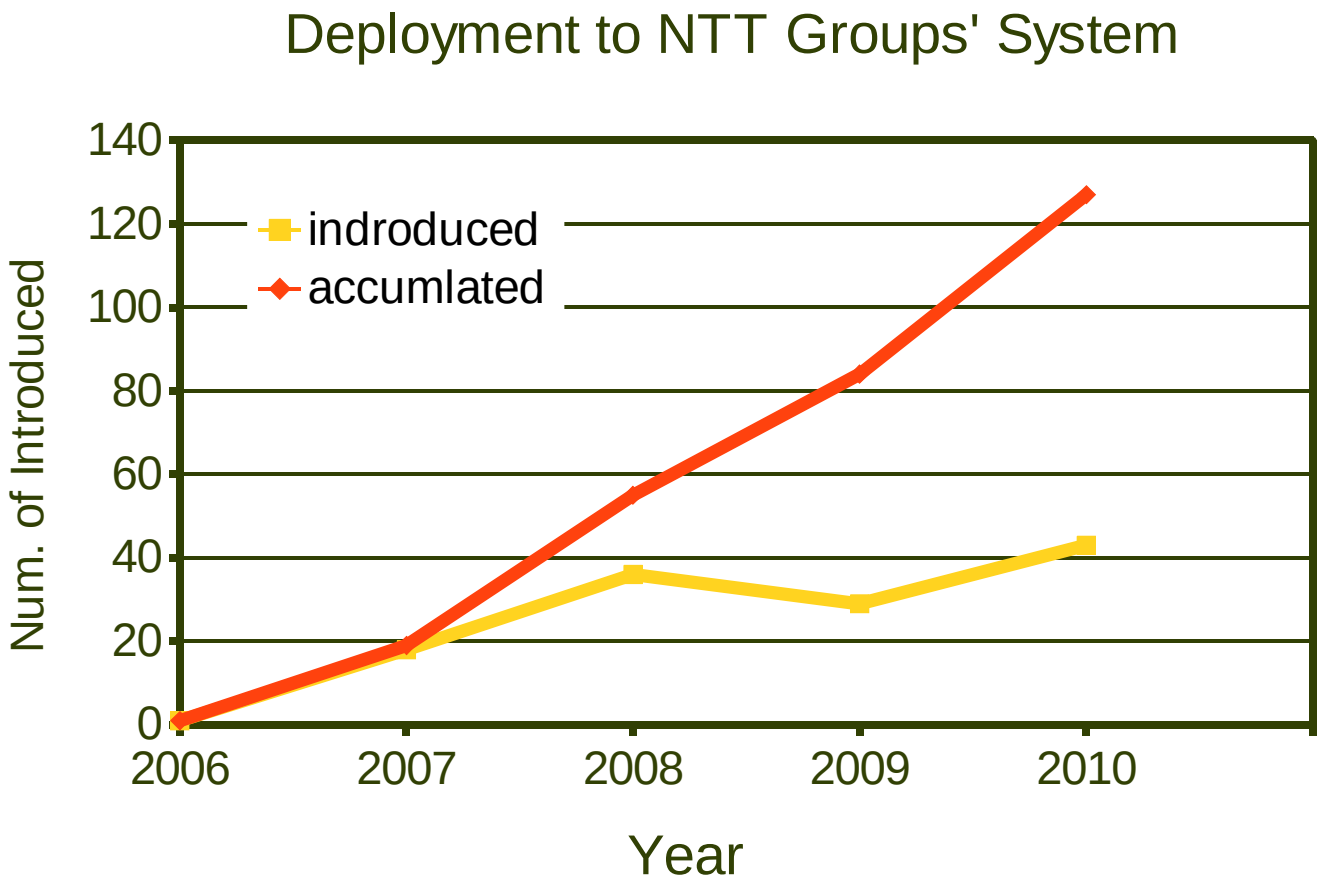
- Target of OSS deployment in NTT in-house systems
  - NTT runs several hundreds systems
  - Survay shows 80% of system can be suitable to deploy PostgreSQL
- Trend of PostgreSQL deployment
  - From small-scale and less available system to large-scale and high available ones



Trend of PostgreSQL Deployment

NTT DATA INTELLILINK CORPORATION

- PostgreSQL was deployed into about 130 systems
- 30-40 systems a year.



[Eyes only] PostgreSQL Application Map

NTT DATA INTELLILINK CORPORATION

- **Sorry, this contents is for eyes only and removed.**

Expectation

NTT DATA INTELLILINK CORPORATION

- Federated DB
  - Large DB system consists of many databases.
- Performance for ‘private cloud'
  - Efficient processing is essential
    - CPU scalable
    - I/O bandwith
- More installations via community
  - More installations improve quality
  - More use cases accelerate introduction

Useful Japanese Sites

NTT DATA INTELLILINK CORPORATION

- Let's Postgres
  - Accumulates useful information of PostgreSQL
    - How-to's
    - Practices
    - Conference reports

    http://lets.postgresql.jp/
- LPI
  - Now have a qualification for Open Source Database (practically PostgreSQL)

    http://www.oss-db.jp/

NTT DATA INTELLILINK CORPORATION

Postgres-XC

What is Postgres-XC?

NTT DATA INTELLILINK CORPORATION

- Short Introductory Video

Overview of Postgres-XC

NTT DATA INTELLILINK CORPORATION

Symmetric PostgreSQL cluster

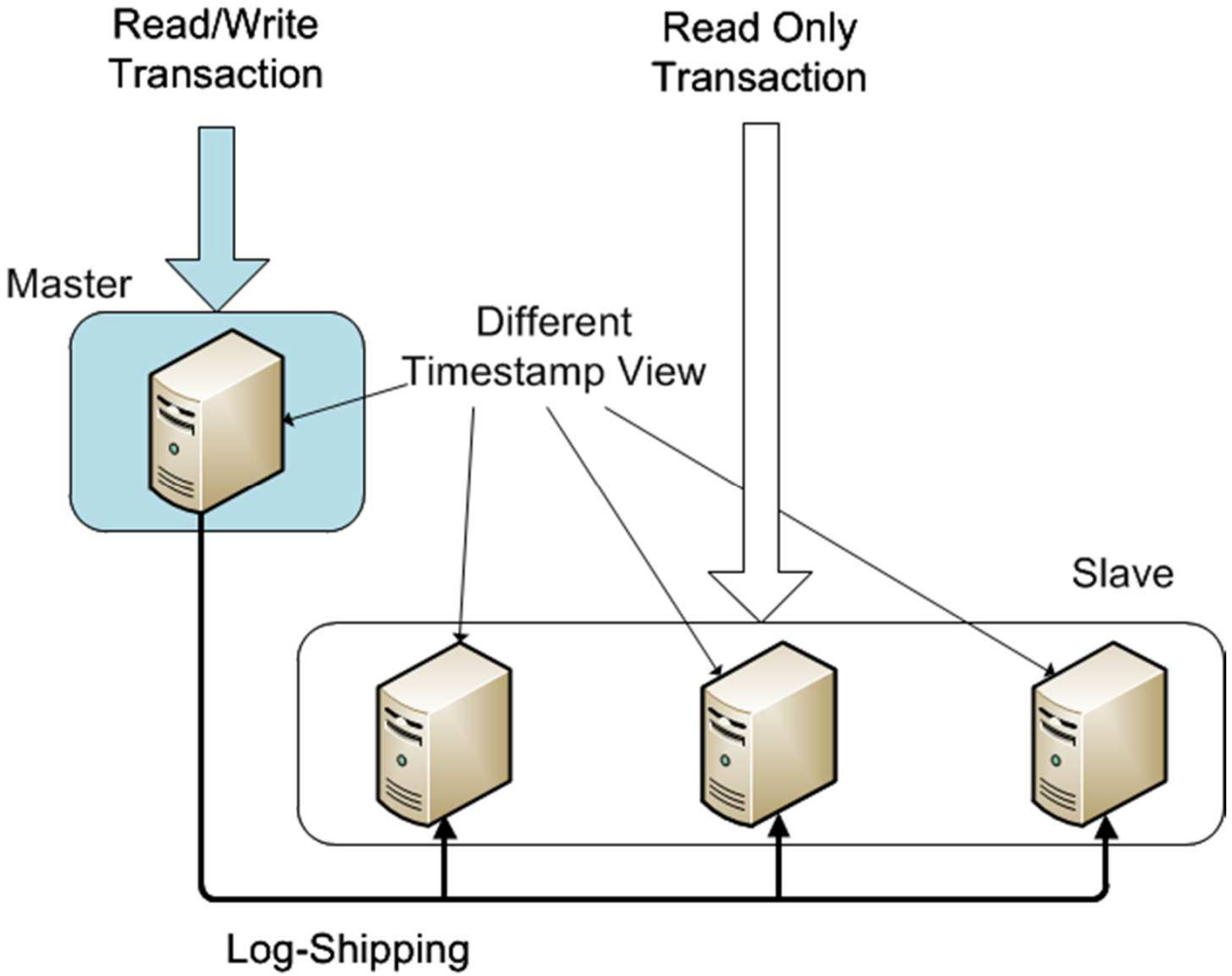
- No Master
- No Slave
  - No READ ONLY slaves
  - Every node can issue both READ/WRITE
- Transparent Transaction Management

Now Version 0.9.6

- Generally available next calendar year

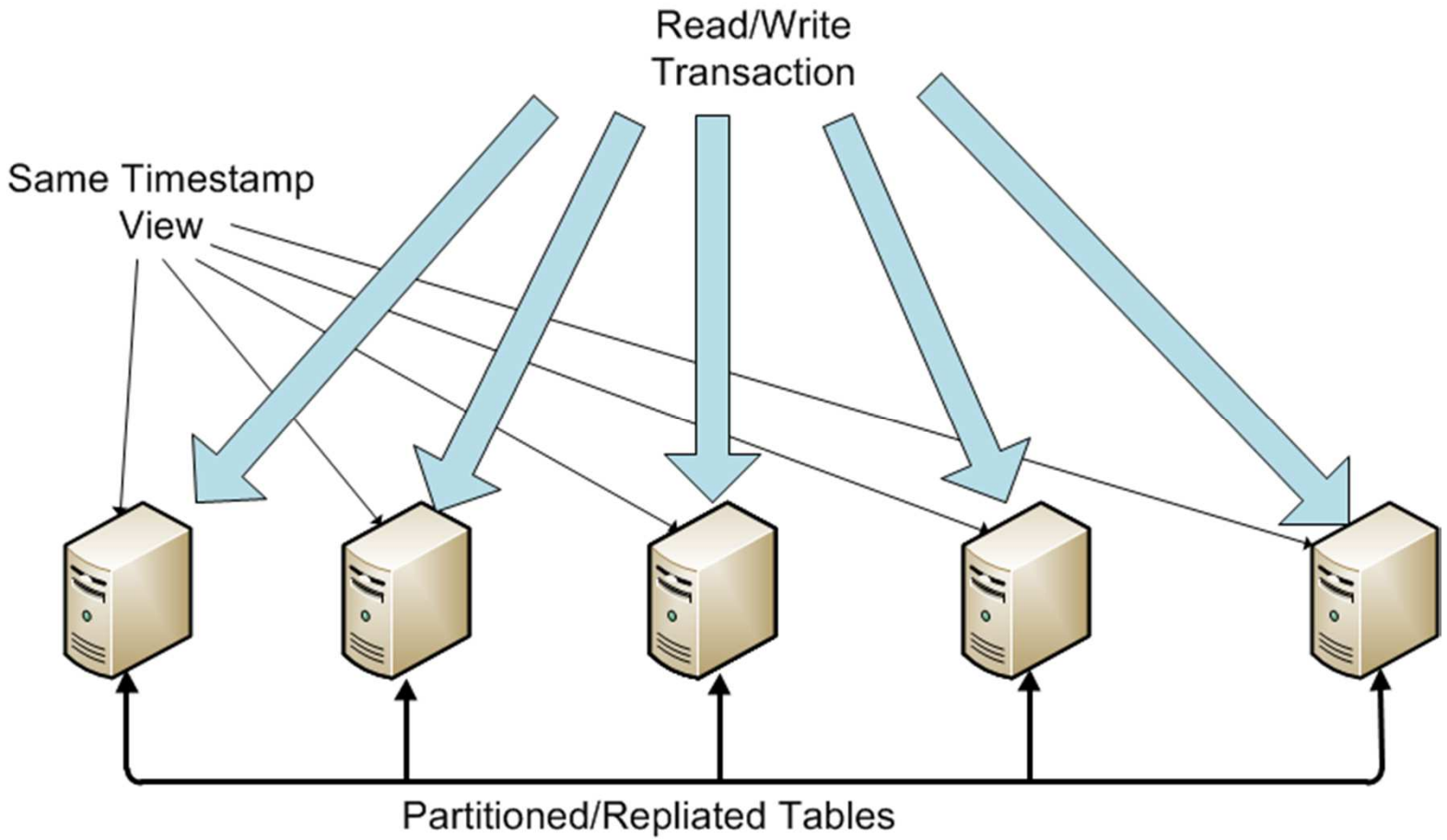
PostgreSQL Master/Slave with Log Shipping

NTT DATA INTELLILINK CORPORATION



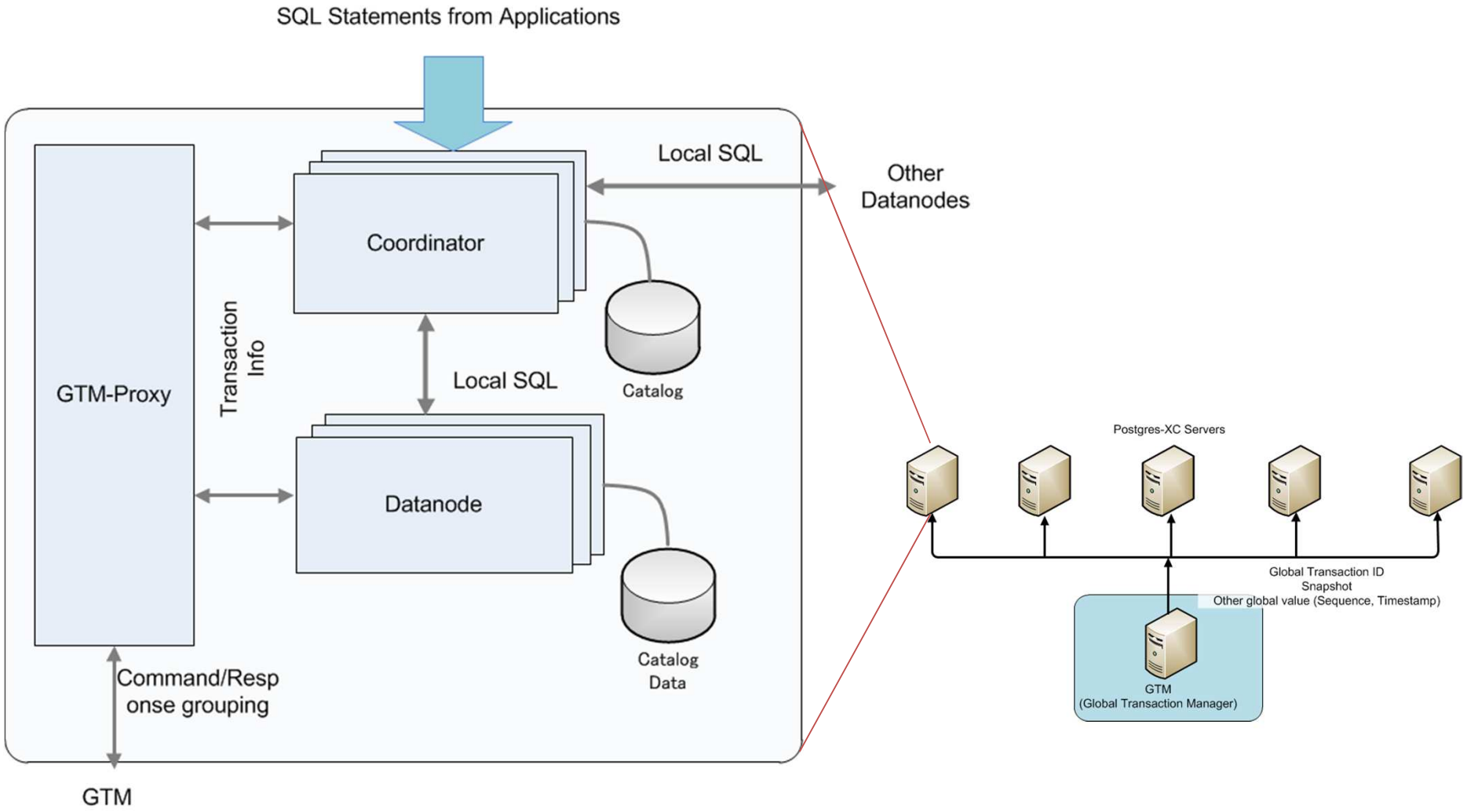
Postgres-XC Symmetric Cluster

NTT DATA INTELLILINK CORPORATION



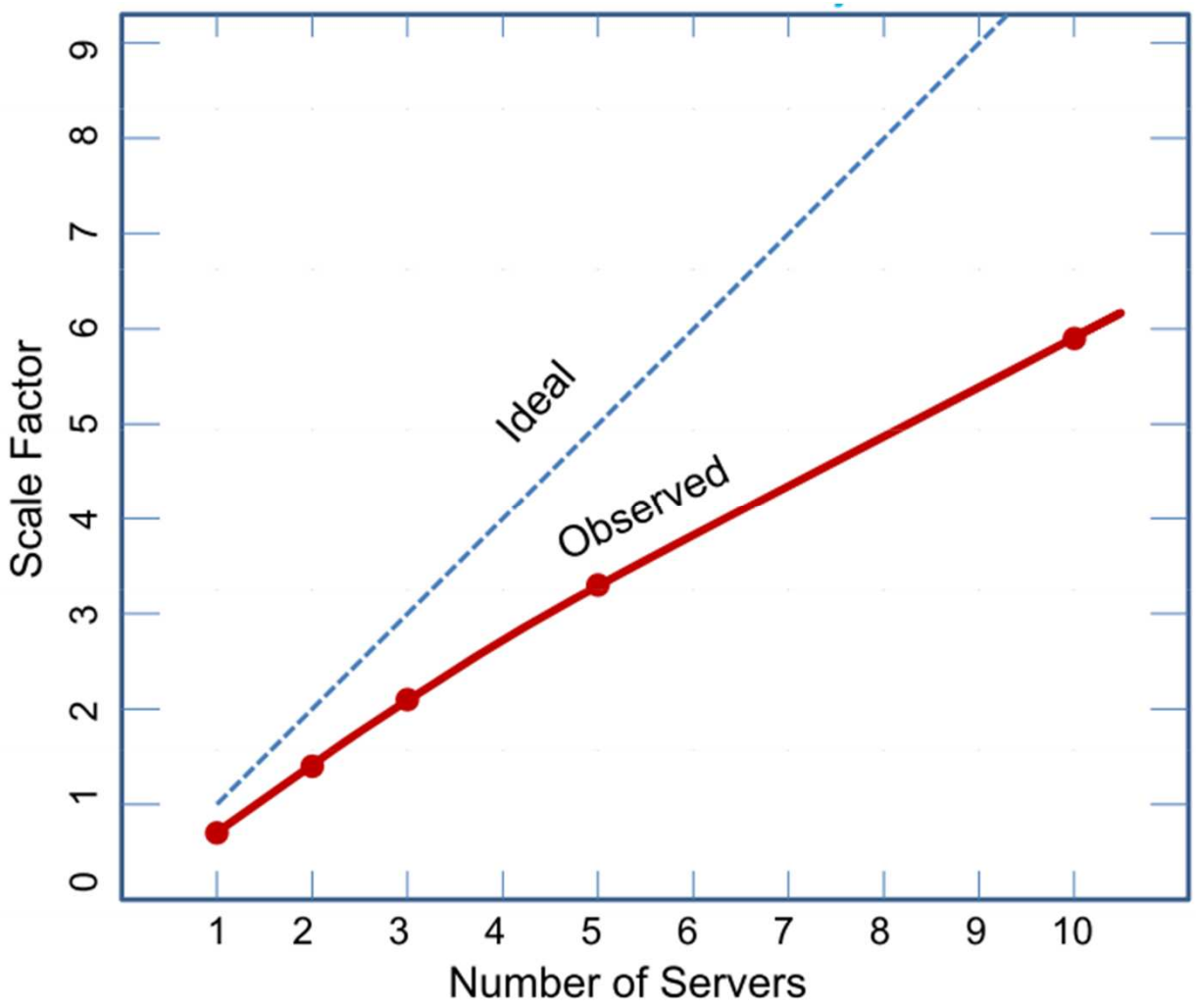
Server Configuration and GTM-Proxy

NTT DATA INTELLILINK CORPORATION



Scalability

NTT DATA INTELLILINK CORPORATION



DBT-1 (Rev)

Current Status

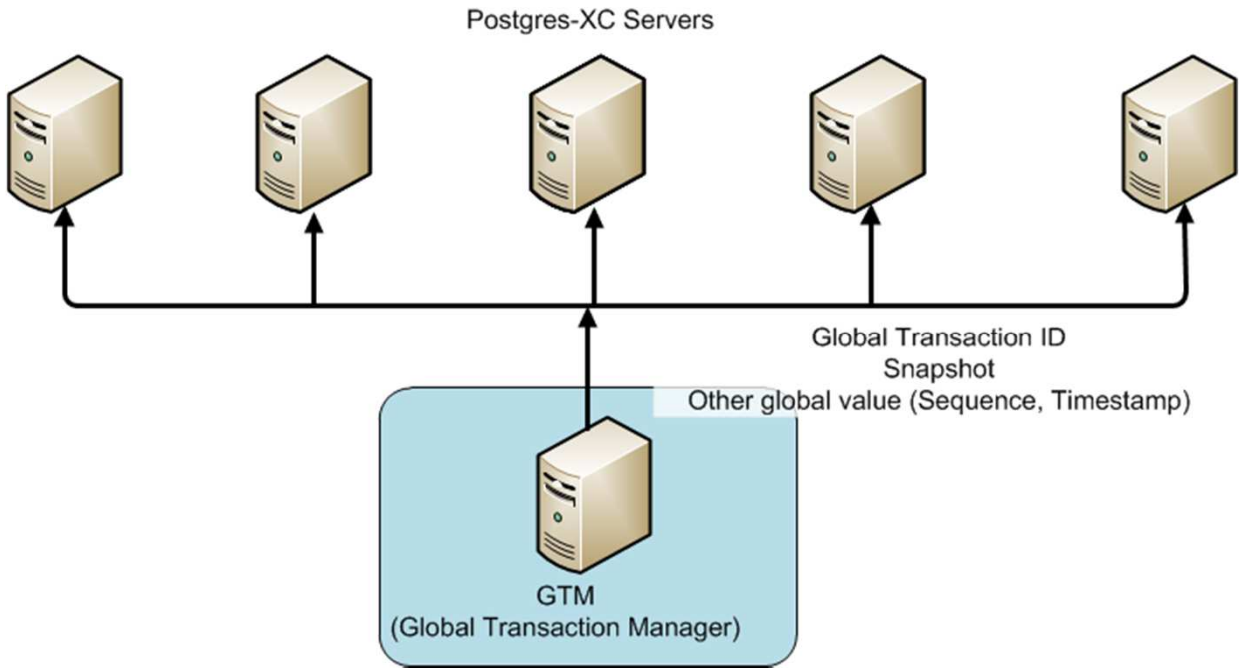
NTT DATA INTELLILINK CORPORATION

- Now V 0.9.6 is available
  - Based upon PostgreSQL 9.1
  - Reference Manual integrated with PostgreSQL reference
- License changed to PostgreSQL license
  - Free to bring outcome back to PostgreSQL

GTM: Key for Transaction Transparency

NTT DATA INTELLILINK CORPORATION

- Consistent Transaction ID (GXID) throughout the system
- Provide global snapshot for consistent visibility from any server



Requirements Since Last Year ...

NTT DATA INTELLILINK CORPORATION

Solution for GTM as SPOF

- GTM Standby

Support same SQL statements as original PostgreSQL

- Functions
- Views
- Cross-node joins
- Role/User/Tablespace
- Transparent DDLs
- Many others

Other High Availability Feature such as

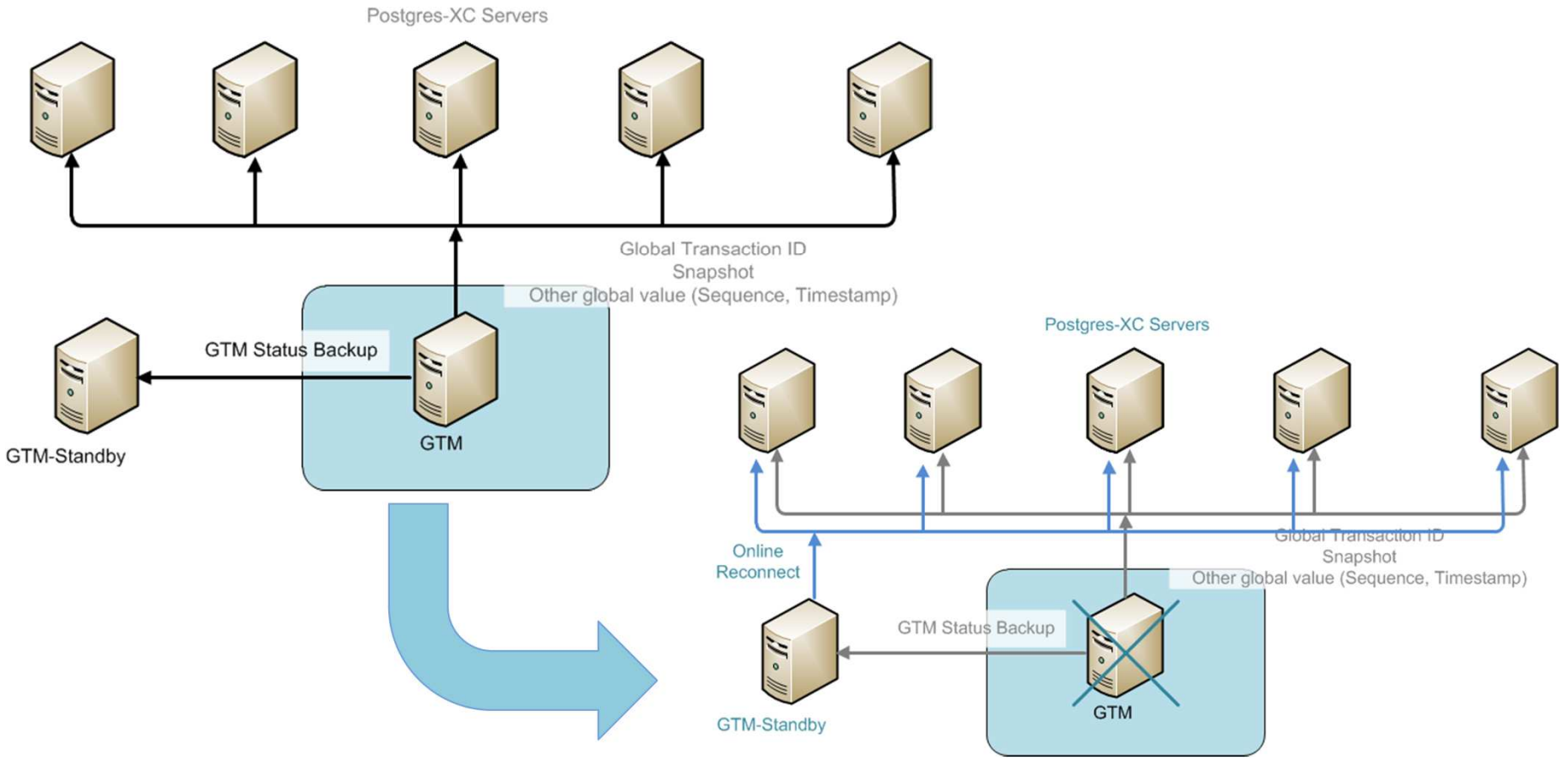
- Data Node Standby
- Consistent Backup and Recovery

Flexible Node Configuration

- On-line addition/Removal

GTM Standby

NTT DATA INTELLILINK CORPORATION



GTM Standby Requirements

NTT DATA INTELLILINK CORPORATION

Online Promote and Reconnect

- Invisible from applications
  - Can be visible from GTM-Proxy
- Transactions should be able to continue to run

GTM-Standby: Current Status

NTT DATA INTELLILINK CORPORATION

- Infrastructure Available: V 0.9.5
- Improvement in progress
  - Connect to GTM at anytime
    - At present, GTM-Standby should be the first to connect to GTM
  - Get rid of any chance of backup information loss
    - Backup first
    - Negotiate the last message at reconnect
  - Performance
    - Backup grouping and decrease response
- Improvement scheduled at the next release

Postgres-XC Statement Extension

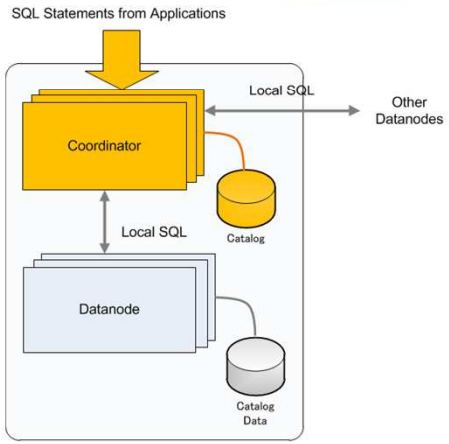
NTT DATA INTELLILINK CORPORATION

Analyze incoming statements

Determine which datanodes are to be involved
- Compose Local Statement for each Datanode

Datanode handle the local statement.
- Result to the coordinator

If needed, materialize the result at the coordinator and calculate the final result.



Postgres-XC Statement Extension

NTT DATA INTELLILINK CORPORATION

Analyze incoming statements

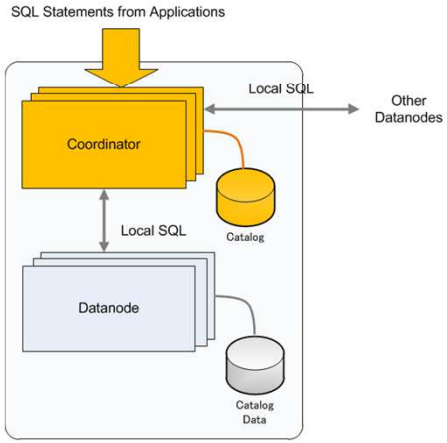
Determine which datanodes are to be involved

- Compose Local Statement for each Datanode

Datanode handle the local statement.

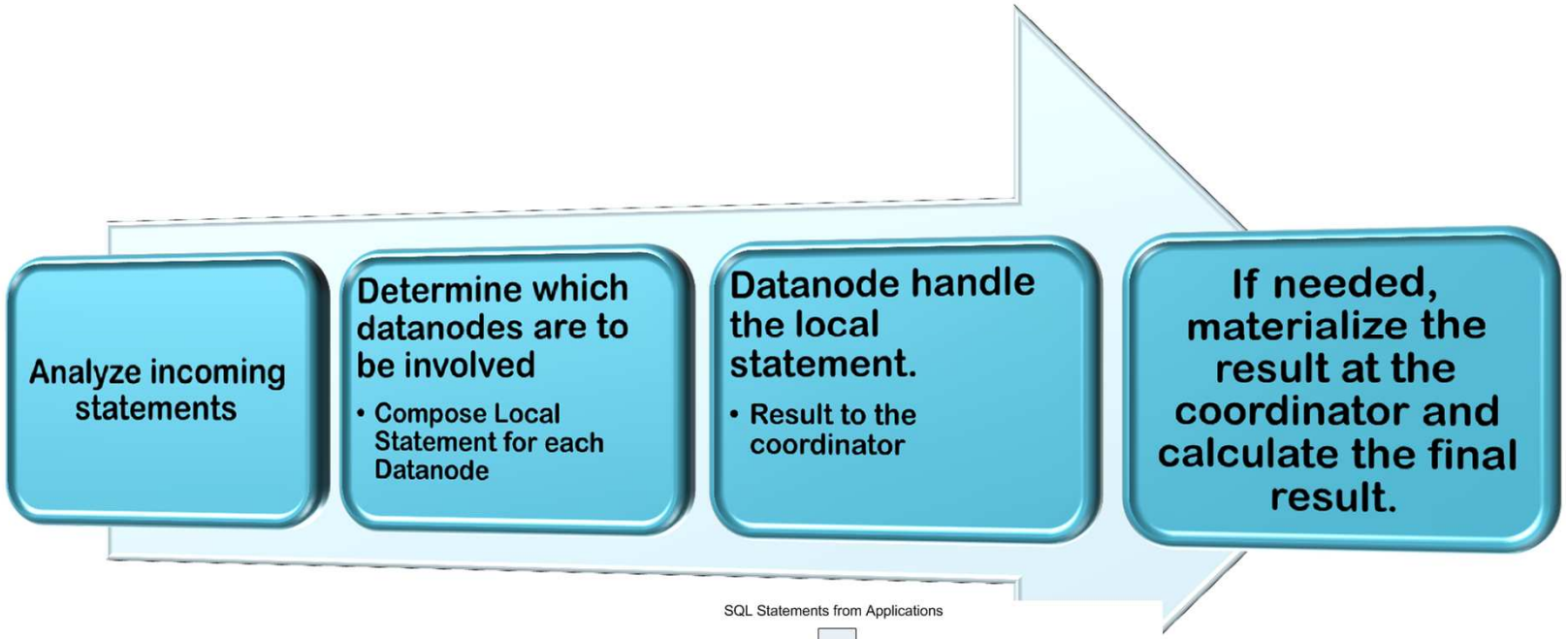
- Result to the coordinator

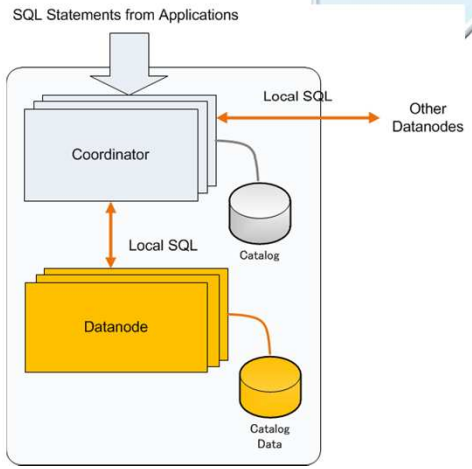
If needed, materialize the result at the coordinator and calculate the final result.



Postgres-XC Statement Extension

NTT DATA INTELLILINK CORPORATION





Postgres-XC Statement Extension

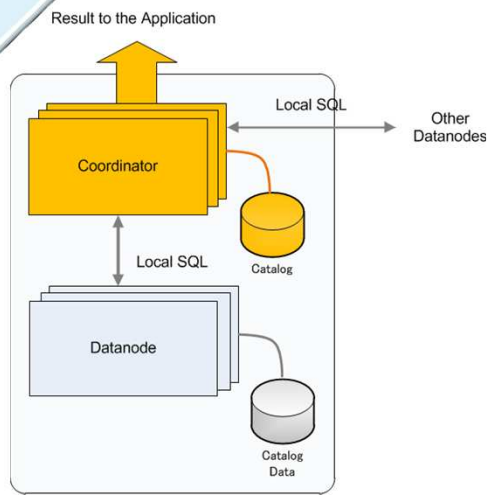
NTT DATA INTELLILINK CORPORATION

Analyze incoming statements

Determine which datanodes are to be involved
- Compose Local Statement for each Datanode

Datanode handle the local statement.
- Result to the coordinator

If needed, materialize the result at the coordinator and calculate the final result.



Optimizing Statements (V 0.9.6)

NTT DATA INTELLILINK CORPORATION

Push-down as many clause as possible

- Join
- WHERE Clause
- Aggregate
- Functions (when used in WHERE clause)
- Column projection

Uses the following information

- If each table is replicated or partitioned
- Partition key
- Partition algorism (Hash/Modulo/Round Robin)

Future Improvement

NTT DATA INTELLILINK CORPORATION

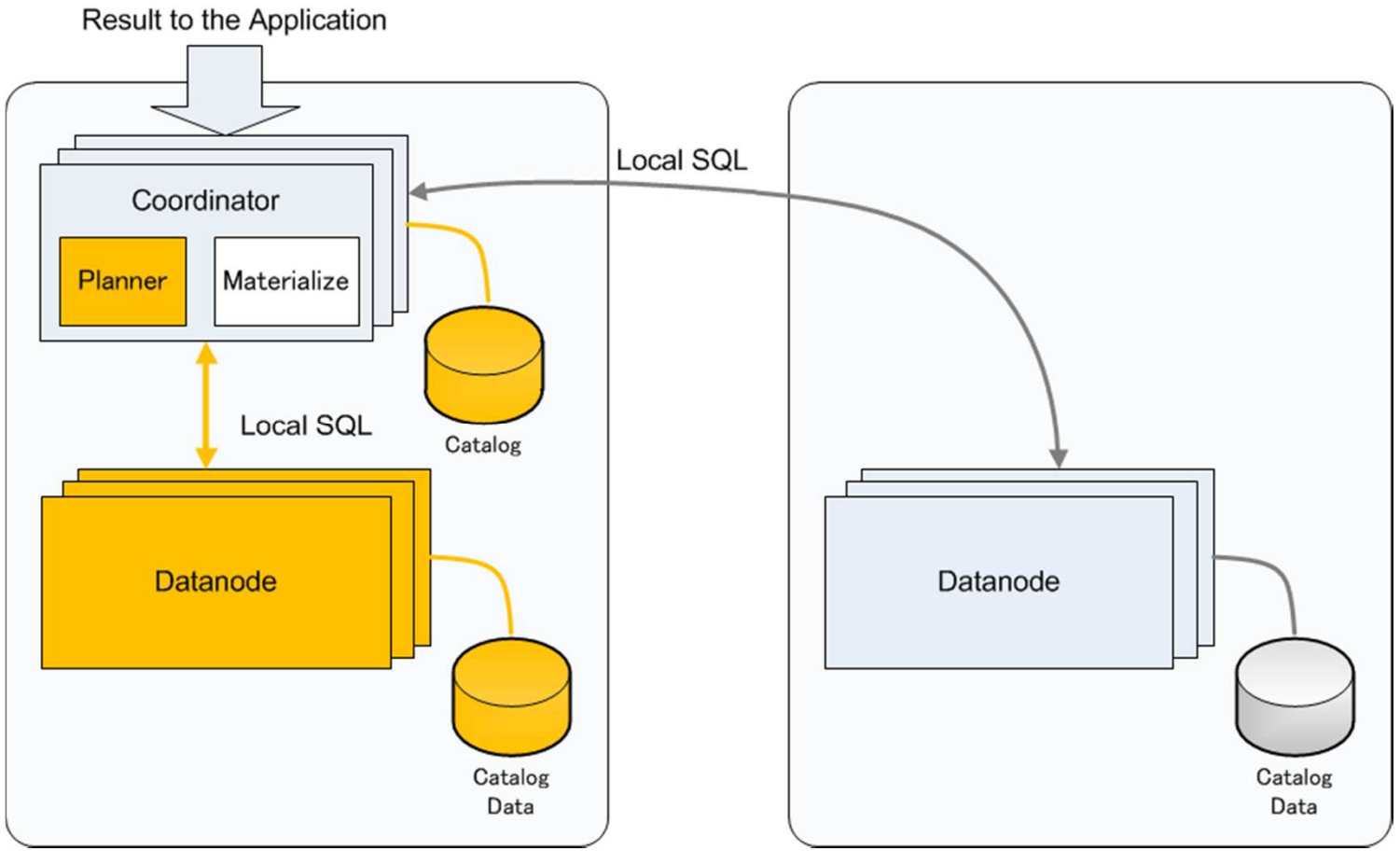
Candidate

- Use statistic info.
- Use Semi-Join to determine joining rows
- Direct join tuple transfer among datanodes
- Much more …

XC Optimization Examples (Join-1)

NTT DATA INTELLILINK CORPORATION

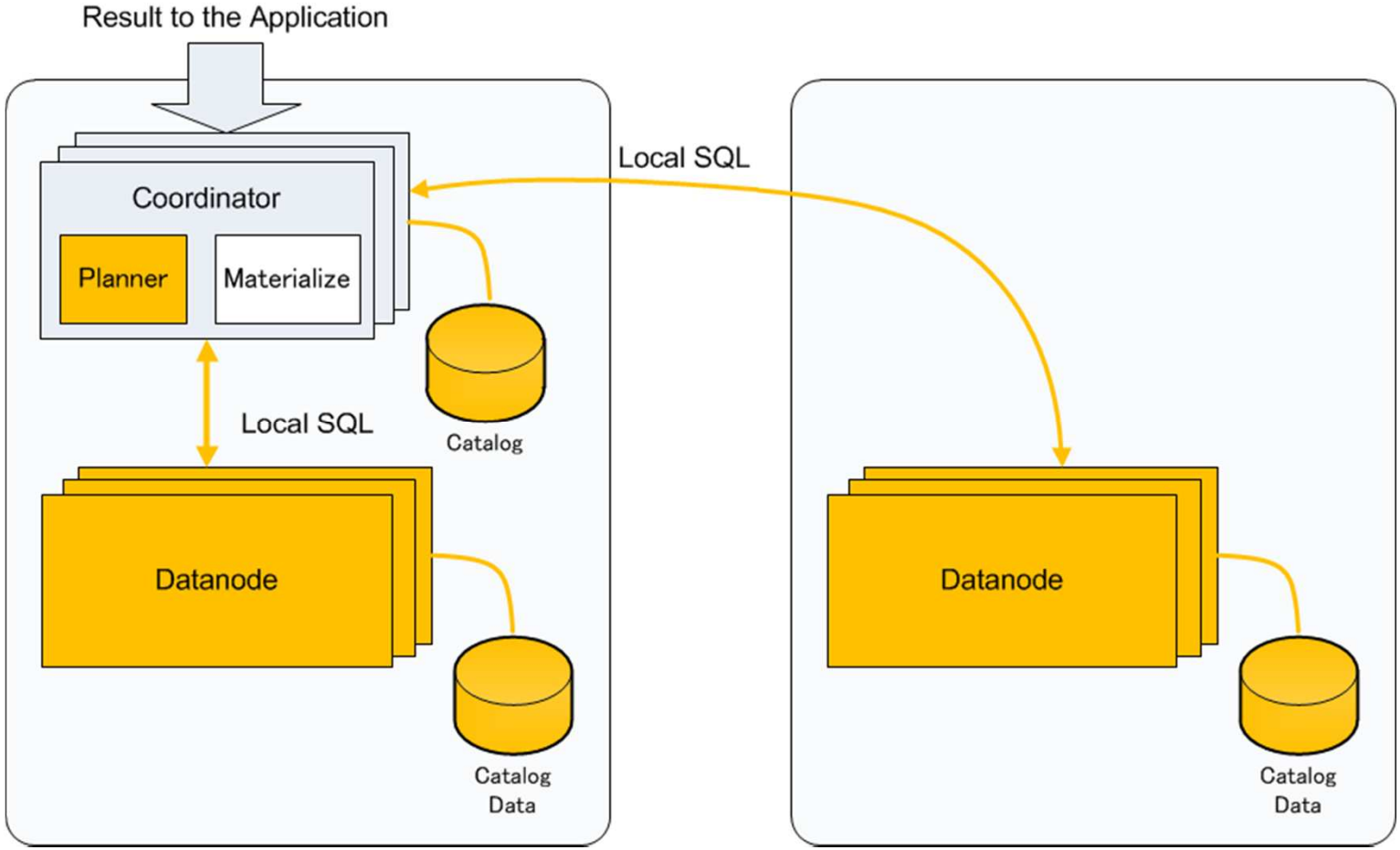
- **Both Tables Are Replicated**



XC Optimization Examples (Join-2)

NTT DATA INTELLILINK CORPORATION

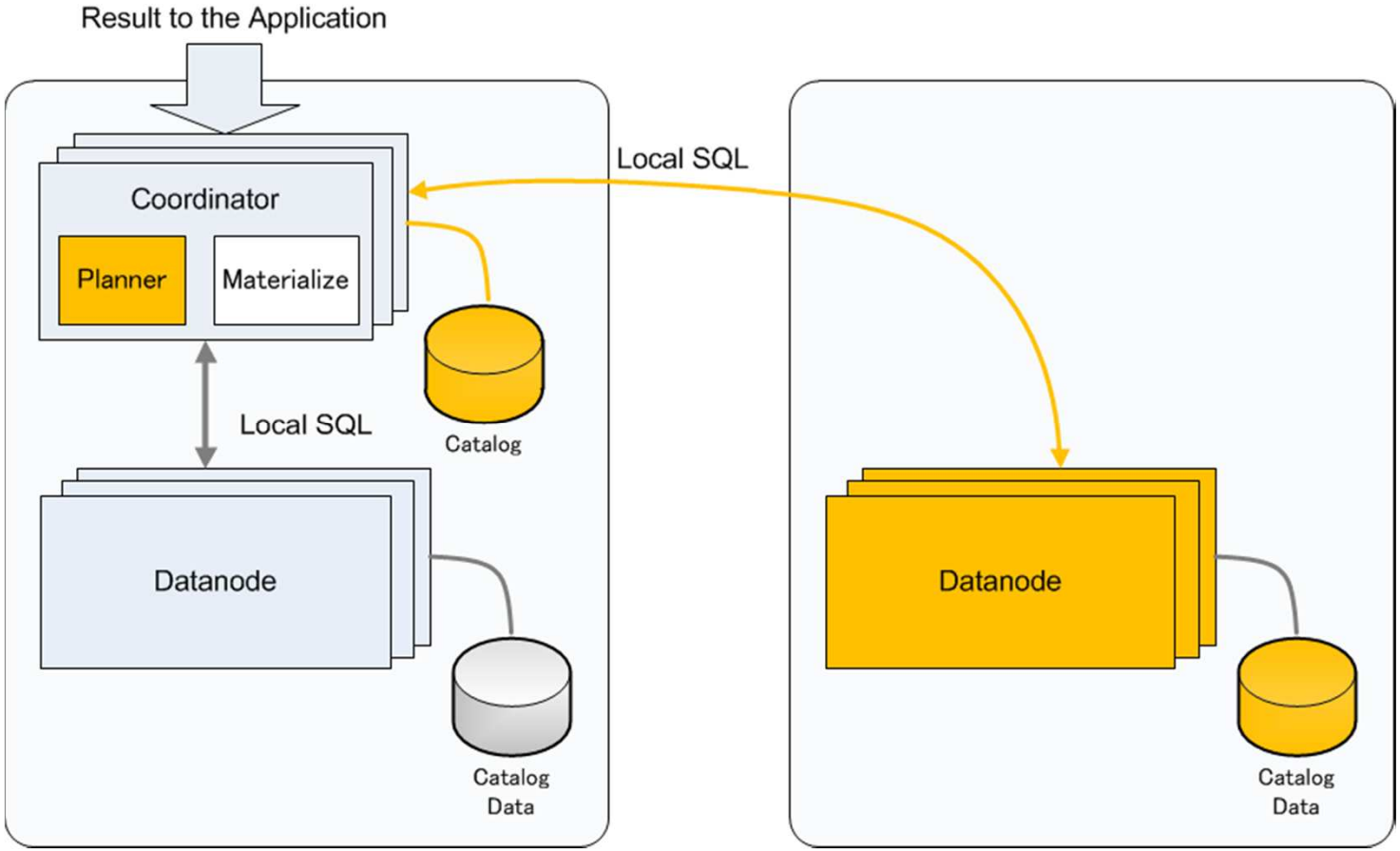
- **Replicated Table and Partitioned Table**



XC Optimization Examples (Join-3)

NTT DATA INTELLILINK CORPORATION

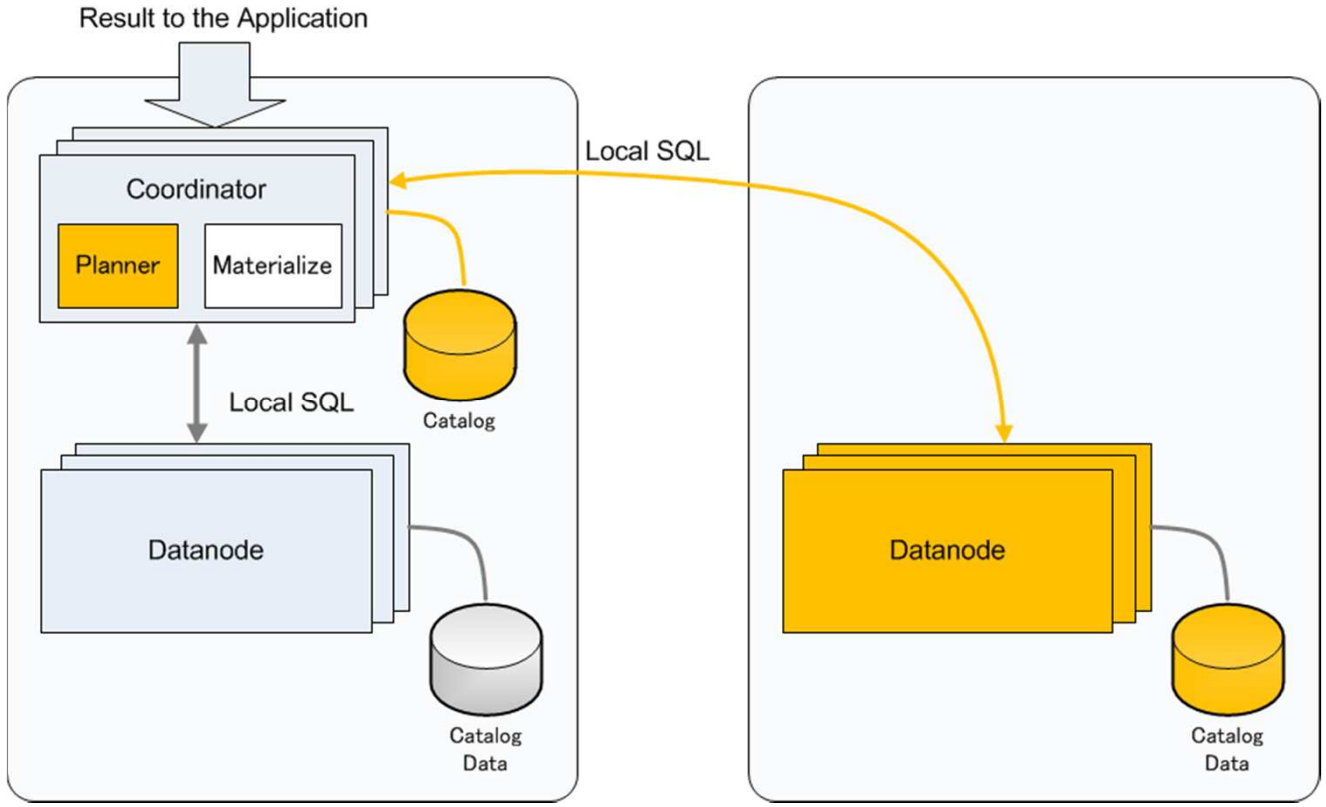
- Replicated Table and Partitioned Table
  - Can determine which datanode to go from WHERE clause



XC Optimization Examples (Join-4)

NTT DATA INTELLILINK CORPORATION

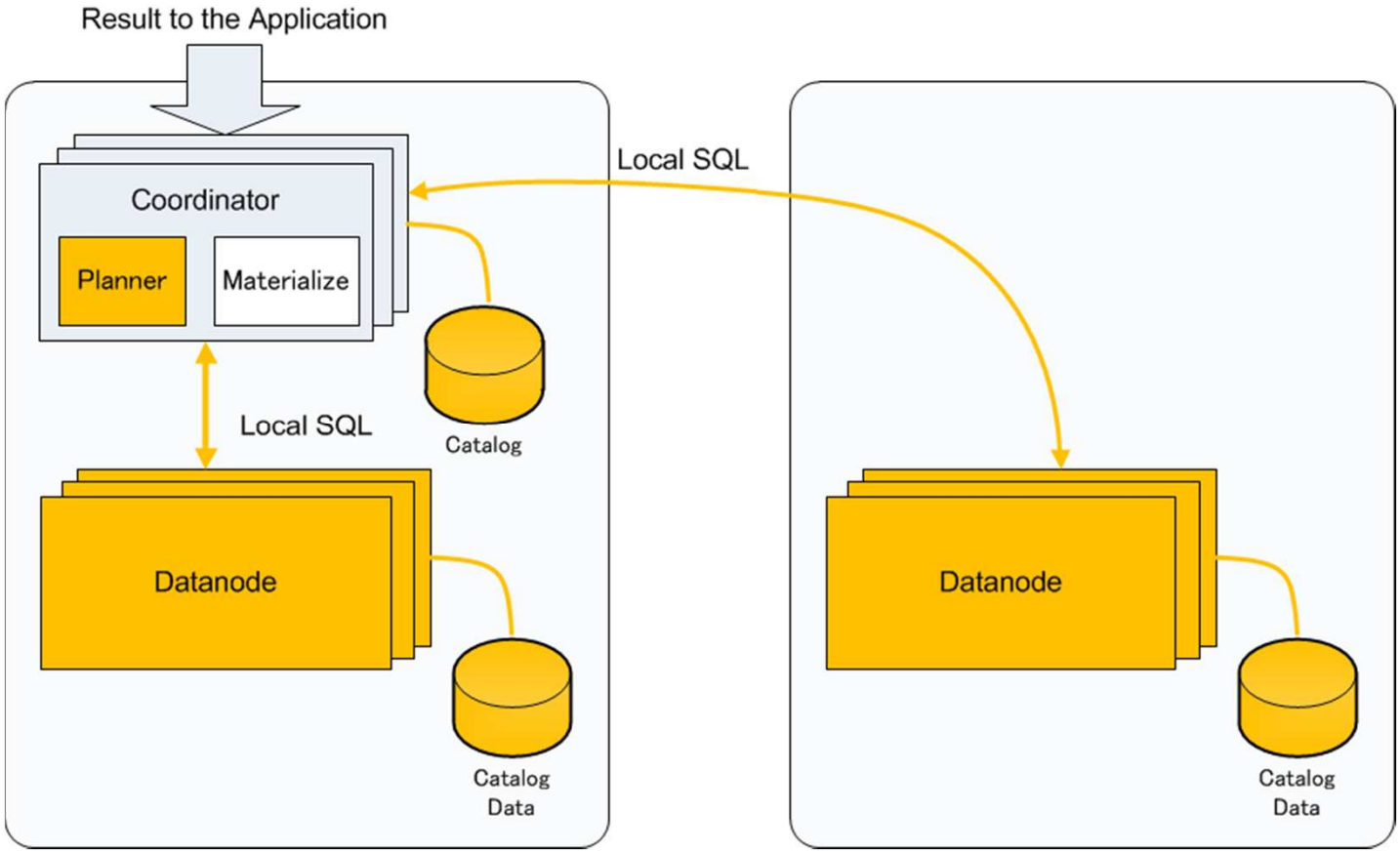
- **Partitioned Table and Partitioned Table**
  - **Both Join columns are distribution (partitioning) column**
  - **Where clause can determine which datanode to go**



XC Optimization Examples (Join-5)

NTT DATA INTELLILINK CORPORATION

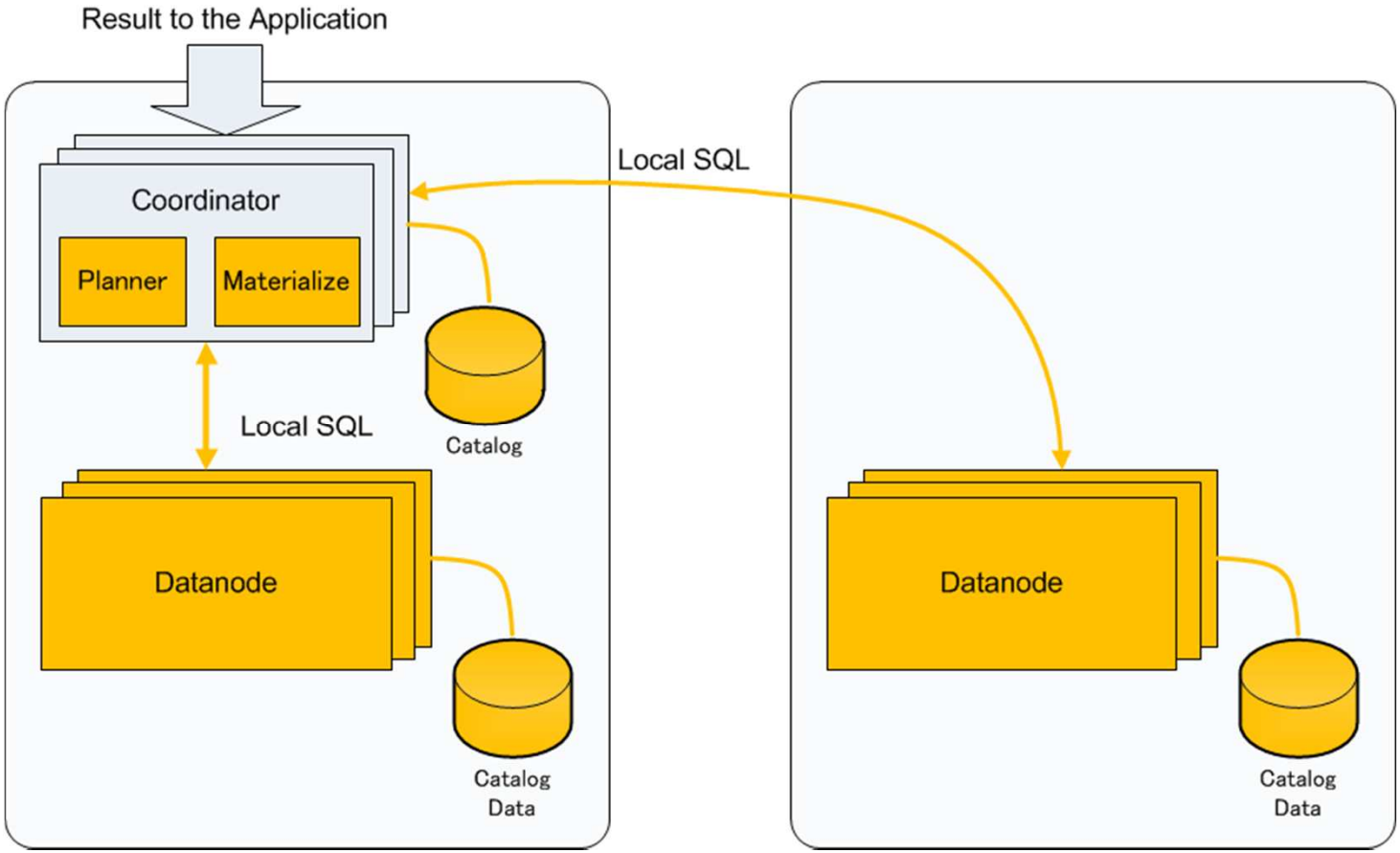
- Partitioned Table and Partitioned Table
  - Both Join columns are distribution (partitioning) column



XC Optimization Examples (Join-6)

NTT DATA INTELLILINK CORPORATION

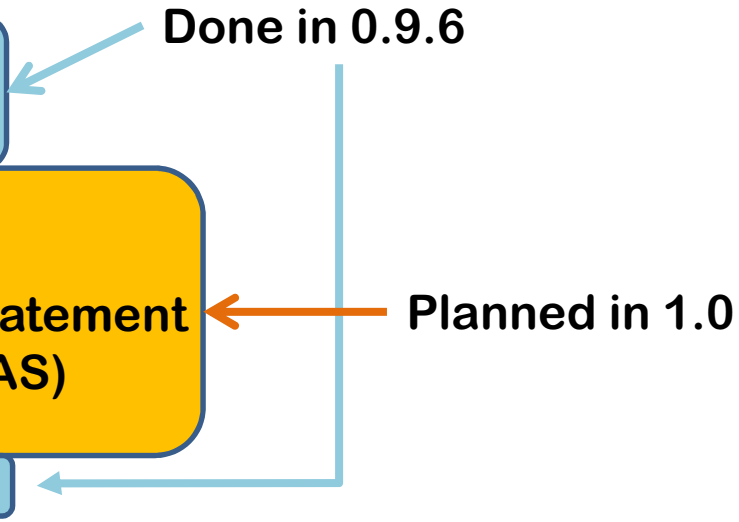
- Partitioned Table and Partitioned Table
  - One of Join columns are not distribution (partitioning) column



XC Statement Handling Summary

NTT DATA INTELLILINK CORPORATION

- Now can handle wide variety of PostgreSQL statement.
- Still in progress
  - HAVING
  - PREPARE, EXECUTE, CURSOR
    - Eliminate restrictions
  - WITH/WITH RECURSIVE
  - General Subqueries
  - Functions with more than one statement
  - SELECT INTO (CREATE TABLE AS)
  - Triggers
  - Temp tables



- Challenges
  - Global constraint
  - More Optimization
  - More Parallelism
- Miscellaneous
  - LISTEN/NOTIFY/UNLISTEN

Backup and Recovery (PITR)Requirement

NTT DATA INTELLILINK CORPORATION

- Transaction status should be consistent
  - Each transaction must be either:
    - Committed in all the involved node
    - Running or aborted in all the involved node
- Write such timing in WALs of all the coordinators and datanodes.
- Application can provide such timing as “BARRIER”
  - CREATE BARRIER *barrier_id*
    - Wait partially-committed-transactions completes commit,
    - Block other transaction’s commit,
    - Write BARRIER record to WALs of all the coordinators/datanodes.
  - When running PITR, specify barrier_id in recovery.conf

Demonstration

NTT DATA INTELLILINK CORPORATION

Further Development Topics/Schedule (1)

NTT DATA INTELLILINK CORPORATION

- Support more variety of statements:
  - CURSOR, TRIGGER
    - By the end of March, 2012
  - SAVEPOINT
    - Beyond April, 2012
  - WITH, WITH RECURSIVE, general functions, general subqueries, SELECT INTO, CREATE TABLE AS
    - By the end of March, 2012

Further Development Topics/Schedule (2)

NTT DATA INTELLILINK CORPORATION

- Datanode high-availability
  - Backup with synchronous streaming replication
    - Synchronous replication needed to maintain data integrity among datanodes.
- Cluster operation
  - Online server addition/removal
- Challenging
  - Global constraint
    - Unique/Reference integrity among partition,
    - Exclusion constraint among partition
  - LOB
- Others needs additional test
  - dblink
  - SQL/MED

Postgres-XC to PostgreSQL

NTT DATA INTELLILINK CORPORATION

- Snapshot cloning
  - Parallel pg_dump
  - Parallel query execution (local/cluster)
- SQL/MED extension
  - Column projection pushdown
  - Join pushdown
  - Function pushdown
- Federation
  - Materialization
  - Cross-node join
  - Cross-node aggregation

Many candidate features.
Need more members for quick actions.

New Developer Wanted

NTT DATA INTELLILINK CORPORATION

- Writing Code
  - New distributed/parallel query handling/optimization
  - HA capabilities
  - Utilities
    - Installation
    - Configuration
    - Operation
  - Bug fixes
  - Back port to PostgreSQL
- Build
  - Creating binaries/distribution packages
- Test
  - Performance evaluation with various benchmarks
  - Finding bugs
  - New feature proposals
- Pilot application
  - Practical applications

Project resources

NTT DATA INTELLILINK CORPORATION

- Development site
  - http://sourceforge.net/projects/postgres-xc/
  - http://sourceforge.net/apps/mediawiki/postgres-xc/
- Project home
  - http://postgres-xc.sourceforge.net/
- Mailing List
  - http://postgres-xc.sourceforge.net/mailinglist.html

Contact us!

NTT DATA INTELLILINK CORPORATION

Thank you very much!
Muinto Obrigado!

Koichi Suzuki

NTT DATA INTELLILINK Corporation

Pacific Marks Tsukishima,1-15-7,
Tsukishima, Chuo-ku,
Tokyo 104-0052, Japan

TEL : +81 3 5843 6800
E-mail : koichi@intellilink.co.jp
koichi.szk@gmail.com
URL : http://www.intellilink.co.jp/ *only in Japanese
http://www.intellilink.co.jp/plan/corporate/fellow_OSS-DB.html