DB2

Overview of Disaster Recovery Options

Sina Weibo @zhoushuoji

DTCC2012

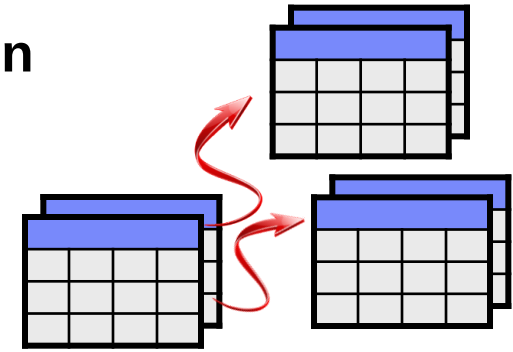
April 16, 2012

Overview of Disaster Recovery Options

- Q Replication
- InfoSphere Change Data Capture (CDC)
- Storage Replication
- Geographically Dispersed pureScale Cluster (GDPC)
- Log Shipping
- HADR
- Comparison of DR Options

Q Replication

- **High-throughput, low latency logical data replication**
  - Distance between sites can be up to thousands of km
- **Asynchronous replication**
  - No “zero data loss” guarantee
- **Includes support for:**
  - Delayed apply
  - Multiple targets
  - Replicating a subset of data
  - Data transformation
- **DR site can be active**
  - Bi-directional replication is supported for updates on both primary and DR sites



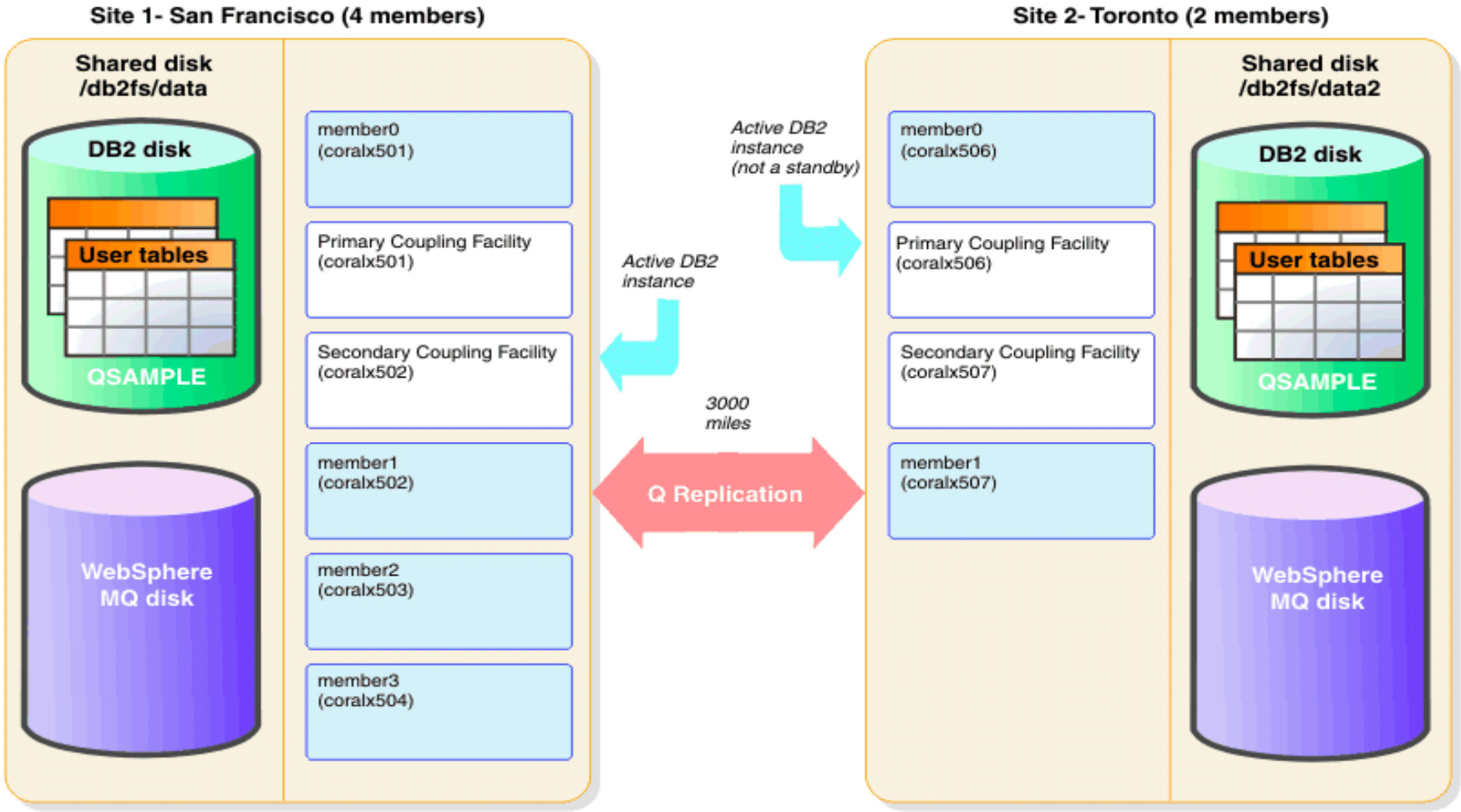
Q Replication

- **Components include Q Capture, Q Apply, and WebSphere MQ**
- **Changes are captured from logs of source database and placed into queues**
- **Highly parallel apply processing applies changes from queues to the target database**
  - Multi-vendor targets supported (e.g. DB2, pureScale, Oracle, etc.)

Q Replication (cont.)

- **Some database applications, certain database constructs, and some types of SQL statements can require special treatment**
  - See the Information Center for details
    http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/topic/com.ibm.swg.im.iis.repl.qrepl.doc/topics/iiyrqplnconstructs.html
- **DB2 to DB2 replication included as part of Advanced Enterprise Server Edition**

Q Replication: Active/Active DR Example

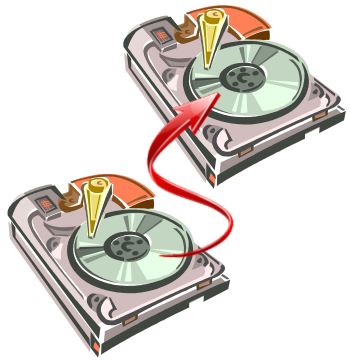


DTCC2012

InfoSphere Change Data Capture (CDC)

- **Similar to Q Replication in that it supports high-throughput, low-latency logical data replication**
  - Does not use MQ queues for data transport
- **Supports bi-directional replication and active/active DR**
- **DB2 can be both source and target of replication**
  - Mix of multi-vendor sources and targets supported

Storage Replication



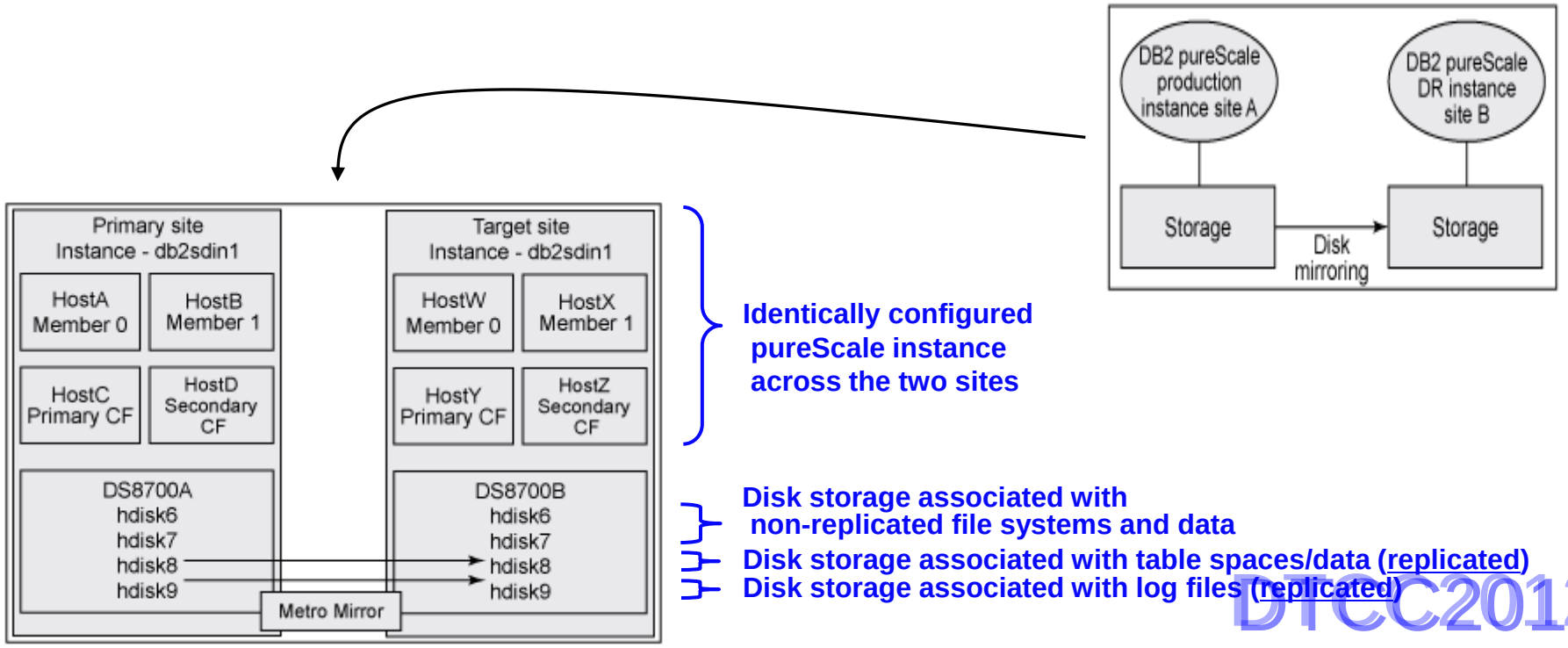
- **Uses remote disk mirroring technology**
  - Maximum distance between sites is typically 100s of kms (for synchronous, 1000s of kms for asynchronous)
  - E.g. IBM Metro Mirror, EMC SRDF
- **Transactions run against primary site only, DR site is passive**
  - If primary site fails, database at DR site can be brought online
  - DR site must be an identical DB2 with matching topology
- **All data and logs must be mirrored to the DR site**
  - Synchronous replication guarantees no data loss
  - Writes are synchronous and therefore ordered, but "consistency groups" are still needed
    - If failure to update one volume, don't want other volumes to get updated (leaving data and logs out of sync)

Storage Replication (cont.)

See developerWorks article on using DS8700 MetroMirror

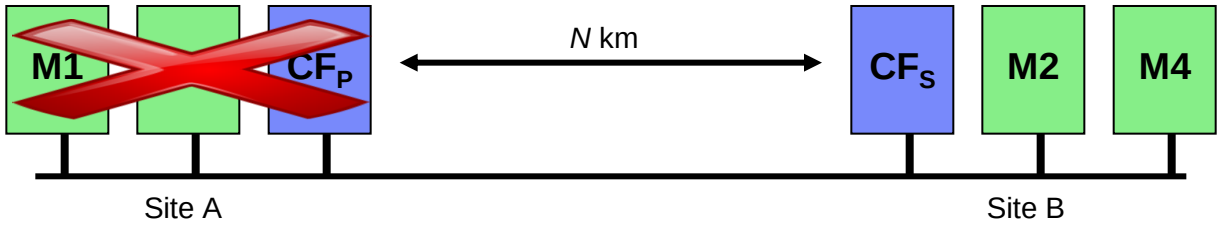
- http://www.ibm.com/developerworks/data/library/techarticle/dm-1005purescalemetromirror

- **Example configuration using DS8700 Metro Mirror**



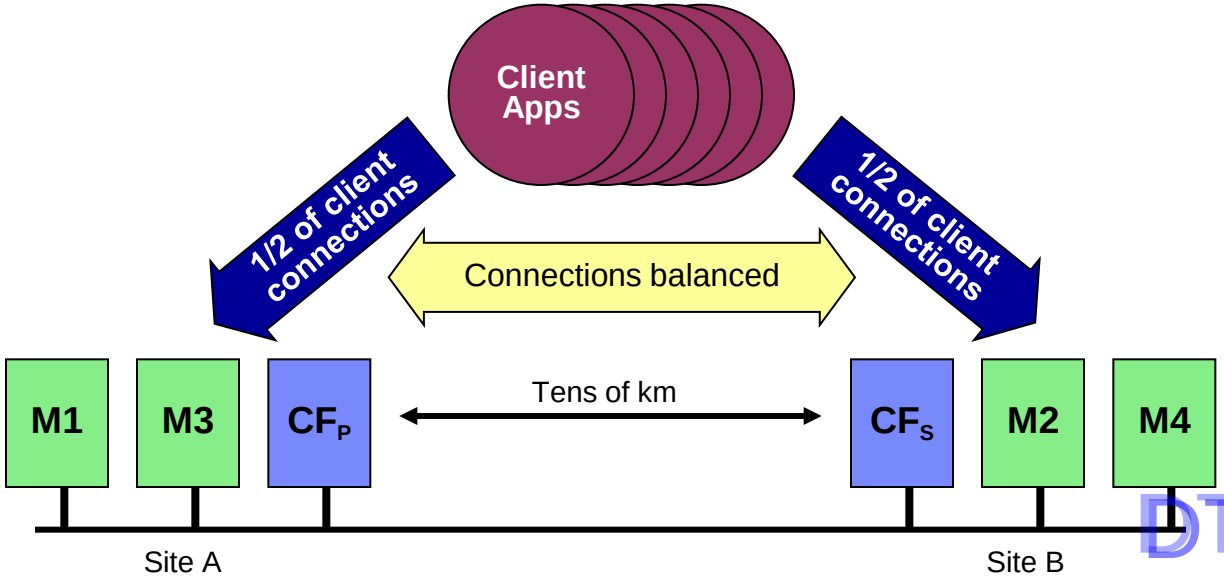
Geographically Dispersed pureScale Clusters (GDPC)

- **A “stretch” or geographically-dispersed pureScale cluster (GDPC) spans two sites A and B at distances of tens of kilometers**
  - Provides active/active access to one or more shared databases across the cluster
  - Enables a level of DR support suitable for many types of disasters



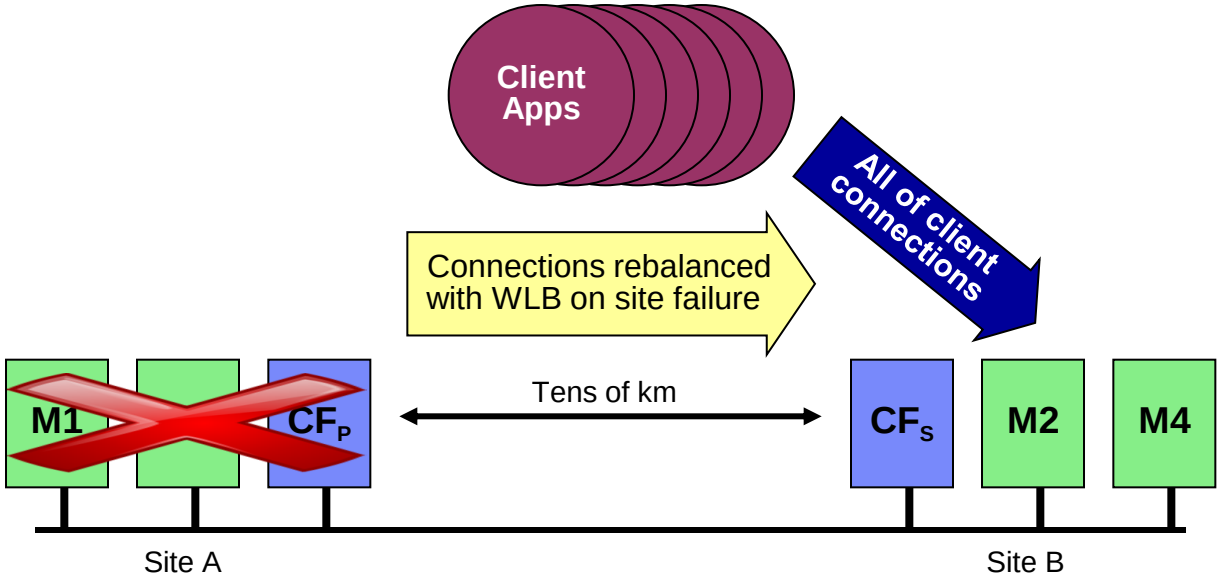
GDPC (cont.)

- **Both sites A and B are active and available for transactions during normal operation**
- **On failures, client connections are automatically redirected to surviving members**
  - Applies to both individual members within sites and total site failure



GDPC Site Failure

- **Handled just like simultaneous failures of member(s) and a CF in a single site pureScale cluster**
  - All client connections go to remaining site
  - One remaining CF is active



Open Systems Extended InfiniBand

- **Typical InfiniBand connectivity reaches at most 10-20 meters**
  - Specialized cables allow up to a few hundred meters
- **IBTA compliant range extenders are compatible with pureScale InfiniBand**
  - We have validated GDPC with Obsidian Research "Longbow" extenders
    http://www.obsidianresearch.com/products/e-series.html
  - Used in pairs, they appear in the network as a 2-port IB switch
  - Convert duplex IB traffic to dark fiber or 10 GbE WAN traffic

Longbow C-103

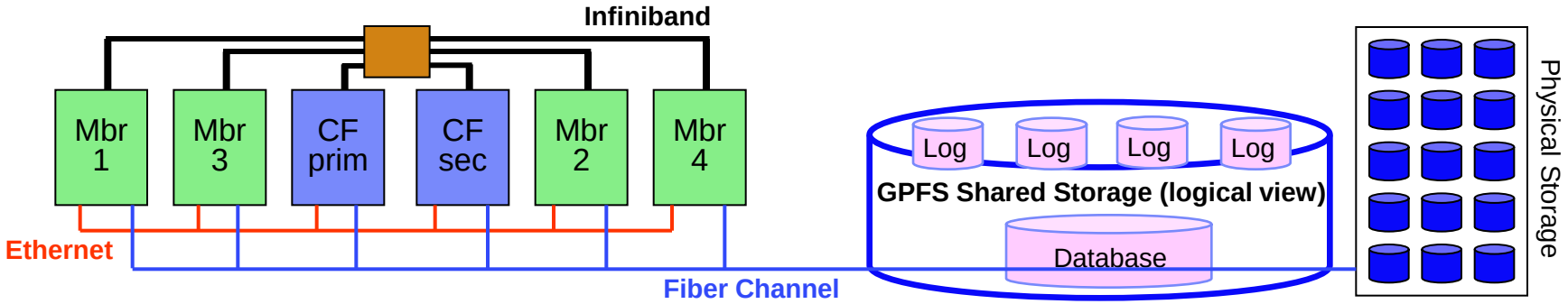
Longbow E-100

Characteristics of a Typical GDPC

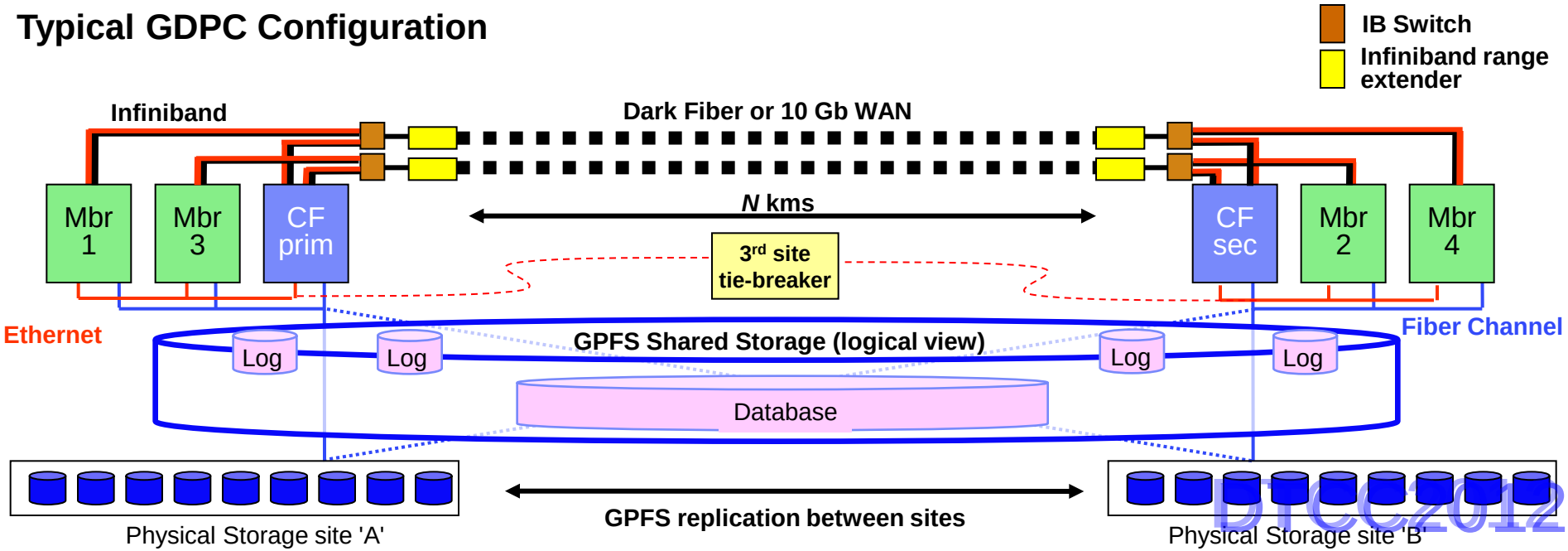
- **Increased message latency of 5 µs / km in glass fiber**
  - For instance, 30 µs CF round-trip @ 3km, 100 µs round-trip at 10km, etc.
  - Greater if repeaters or "slow" WAN are used
  - Can have a negative impact on cluster performance
- **Workloads with a greater portion of read activity (SELECTs) versus writes tend to see lower impact due to distance**
  - GDPC best suited for higher read content workloads (i.e. 80% or more read activity)
  - Impact of read/write ratio grows with distance between sites
- **Workload balancing (WLB) and automatic client reroute (ACR) used to reroute client connections in the event of an outage**
- **Utilizes GPFS synchronous replication to maintain single file system image across the distributed cluster**
- **No SCSI-3 P/R support**
  - No functional impact but adds an estimated one minute of recovery time in the event of a member hardware failure

Typical GDPC Configuration

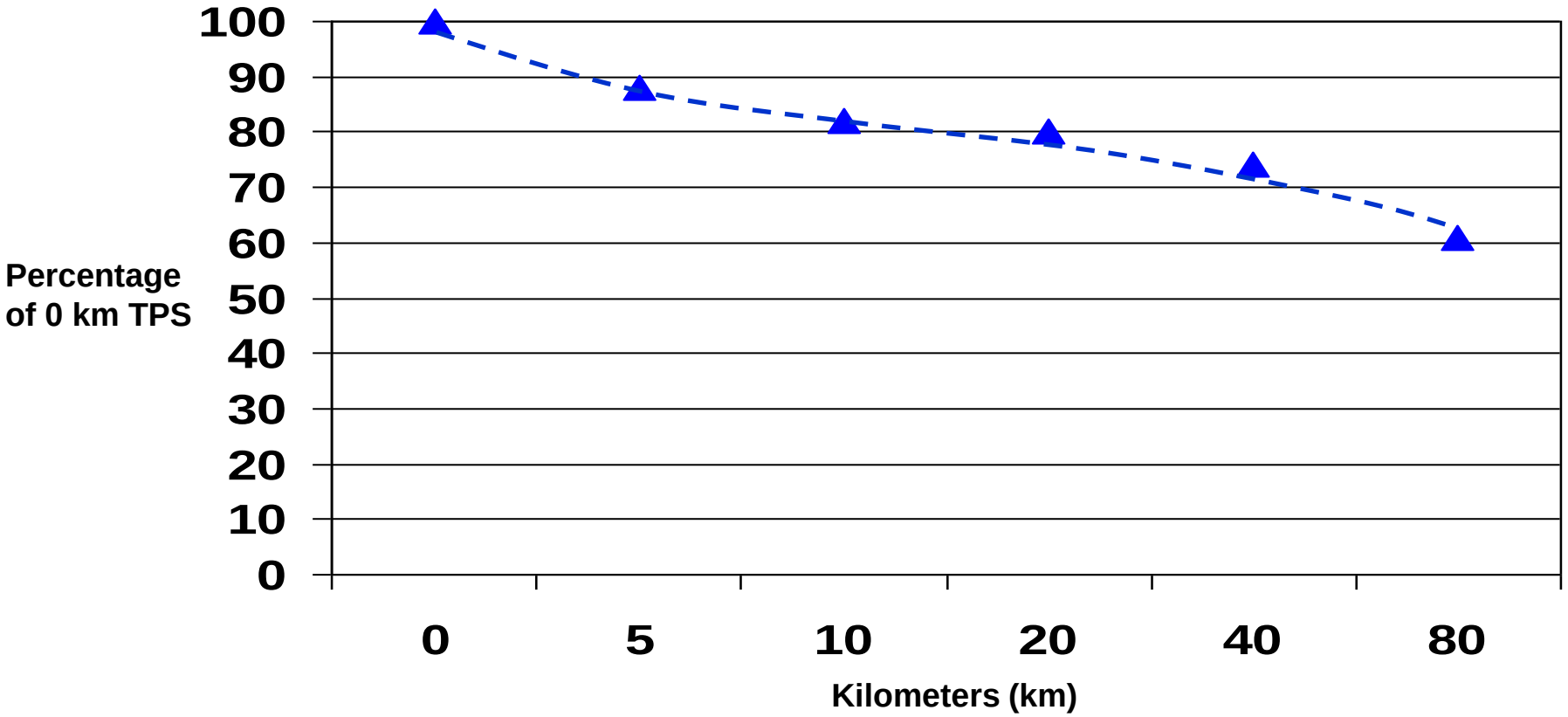
Typical Single-Site Configuration



Typical GDPC Configuration

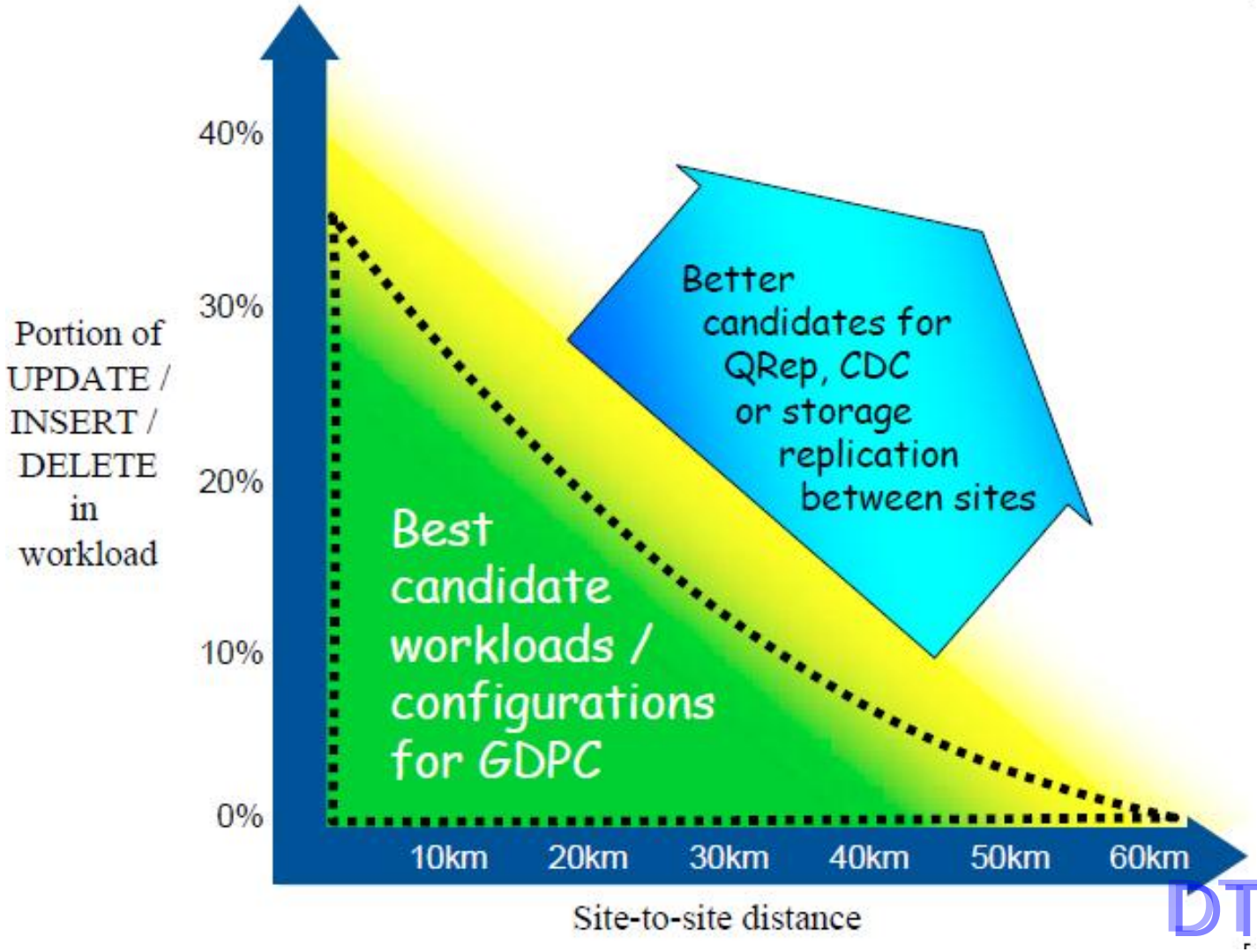


Performance Sensitivity to Distance: Sample Data



Internal OLTP workload with 70% reads 30% writes

Suitability of GDPC



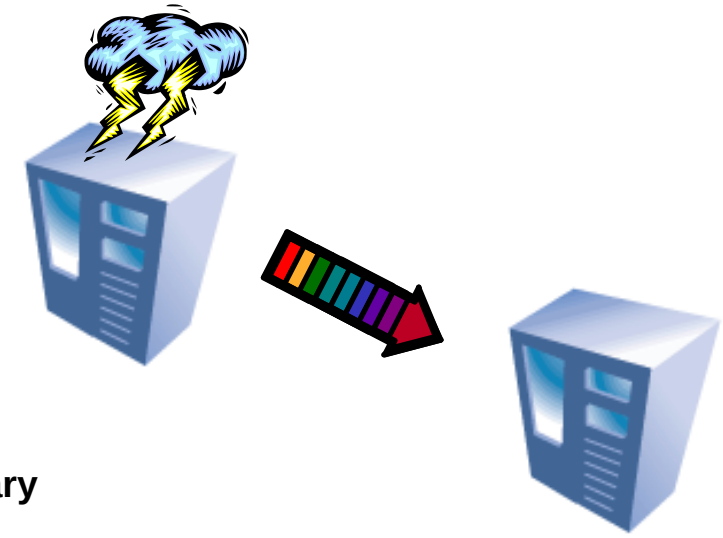
Log Shipping

- **“Home grown” (i.e. user managed) active/passive DR solution**
- **Database on standby system is kept in a perpetual “rollforward in progress” state**
  - Roll forward command is executed repeatedly as log files become available
    - Can choose to incorporate a time delay between the primary and standby
  - `STOP` option is used to bring it out of roll forward state if primary fails
- **Use log files from the archive location and/or use scripts to manage the transfer of log files to the standby site**
  - Recommended that archive location should be geographically separate from the primary site
  - Consider using the `ARCHIVE LOG` command if logs are being filled too slowly
- **Two ways to initialize a standby**
  - Restore of a backup image taken on the primary
  - Using the `db2inidb` command with the `STANDBY` option against a split mirror copy of the primary
- **Operations that are not logged will not be replayed on the standby database**

HADR Main Goals of the Design

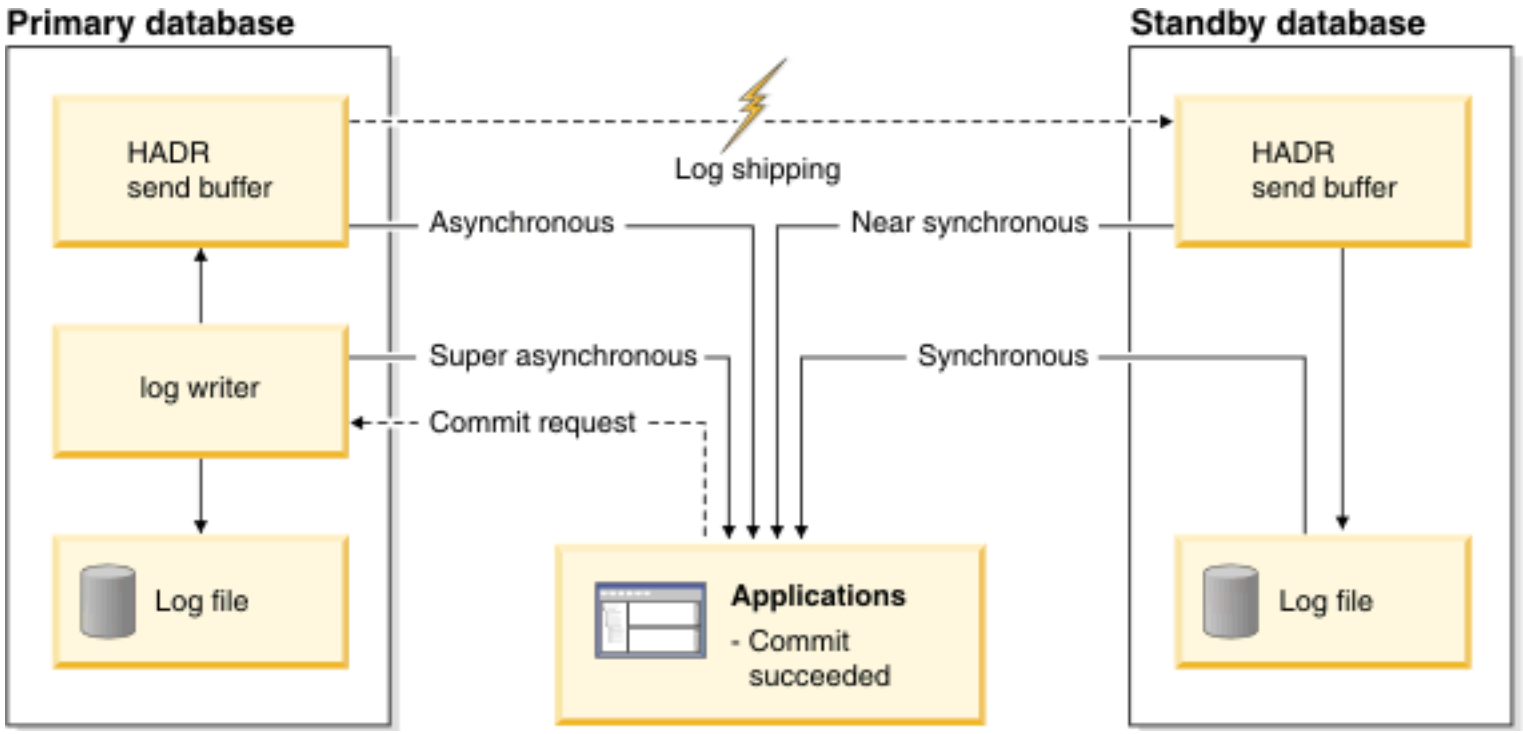
- Ultra-fast failover
- Easy administration
- Handling of site failures
- Negligible impact on performance
- Configurable degree of consistency
- Protection against errant transactions
- Software upgrades without interruption
- Very easy integration with HA-software
- Eventually, no need for HA-software at all
- Transparent failover and failback for applications (combined with “client re-route”)

Basic Principles of HADR



- **Two active machines**
  - **Primary**
    - **Processes transactions**
    - **Ships log entries to the other machine**
  - **Standby**
    - **Cloned from the primary**
    - **Receives and stores log entries from the primary**
    - **Re-applies the transactions**
- **If the primary fails, the standby can take over the transactional workload**
  - **The standby becomes the new primary**
- **If the failed machine becomes available again, it can be resynchronized**
  - **The old primary becomes the new standby**

HADR synchronization mode



Comparison of DR Options

	Synchronous Storage Replication	GDPC	Q Replication / CDC	Log Shipping	HADR
Active/active DR	No	Yes	Yes	No	No
"No transaction loss" guarantee	Yes	Yes	No	No	Yes
Delayed apply	No	No	Yes	Yes	Yes
Multiple targets	No	No	Yes	Yes	Yes*
Maximum distance between sites	100s km	10s km	1000s km (global)	1000s km (global)	N/A

Summary of Disaster Recovery Options in DB2

- Q Replication
- InfoSphere Change Data Capture (CDC)
- Storage Replication
- Geographically Dispersed pureScale Cluster (GDPC)
- Log Shipping
- HADR