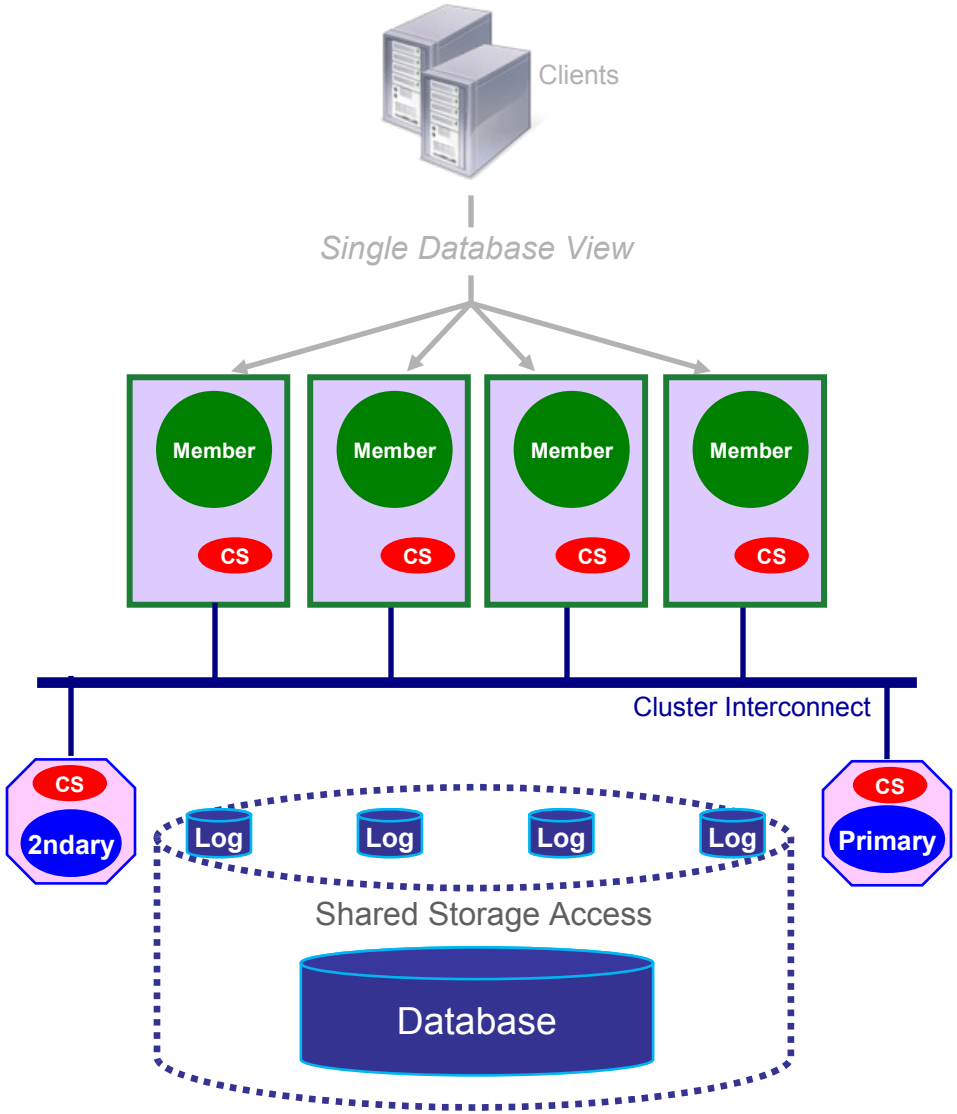
DB2 pureScale stretch cluster
- Long distance call using pureScale

Chat with the Lab for GCG

Xun Xue, IBM Toronto Lab

IBM Software
Information On Demand 2011

DB2 pureScale : Technology Review



Clients connect anywhere,… … see single database

- Clients connect into any member
- Automatic load balancing and client reroute may change underlying physical member to which client is connected

DB2 engine runs on several host computers

- Co-operate with each other to provide coherent access to the database from any member

Integrated cluster services

- Failure detection, recovery automation, cluster file system
- In partnership with STG (GPFS,RSCT) and Tivoli (SA MP)

Low latency, high speed interconnect

- Special optimizations provide significant advantages on RDMA-capable interconnects like Infiniband

Cluster Caching Facility (CF)

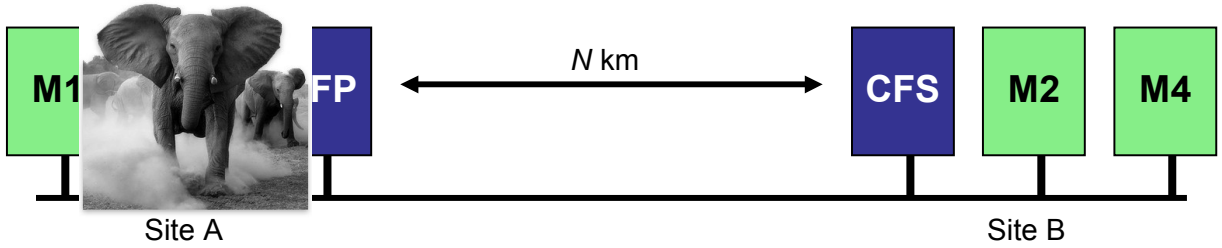
- Efficient global locking and buffer management
- Synchronous duplexing to secondary ensures availability

Data sharing architecture

- Shared access to database
- Members write to their own logs
- Logs accessible from another host (used during recovery)

Active/Active Disaster Recovery via "Stretch Cluster"

- A 'stretch' or geographically-dispersed pureScale cluster (GDPC) spans two sites A & B at distances of tens of km
  - Goal: provide active / active access to one or more shared databases across the cluster
  - Enables a level of DR support suitable for many types of disaster

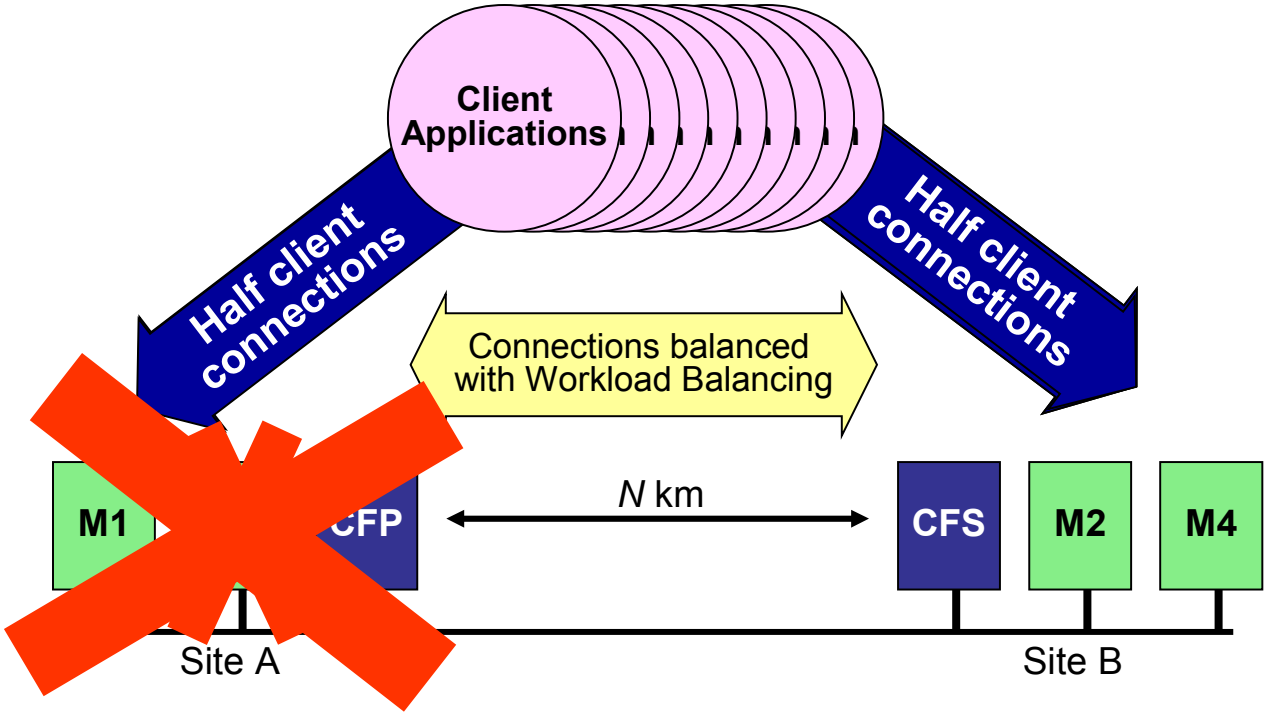


- Inspired by DB2/z Geographically Dispersed Parallel Sysplex (GDPS)

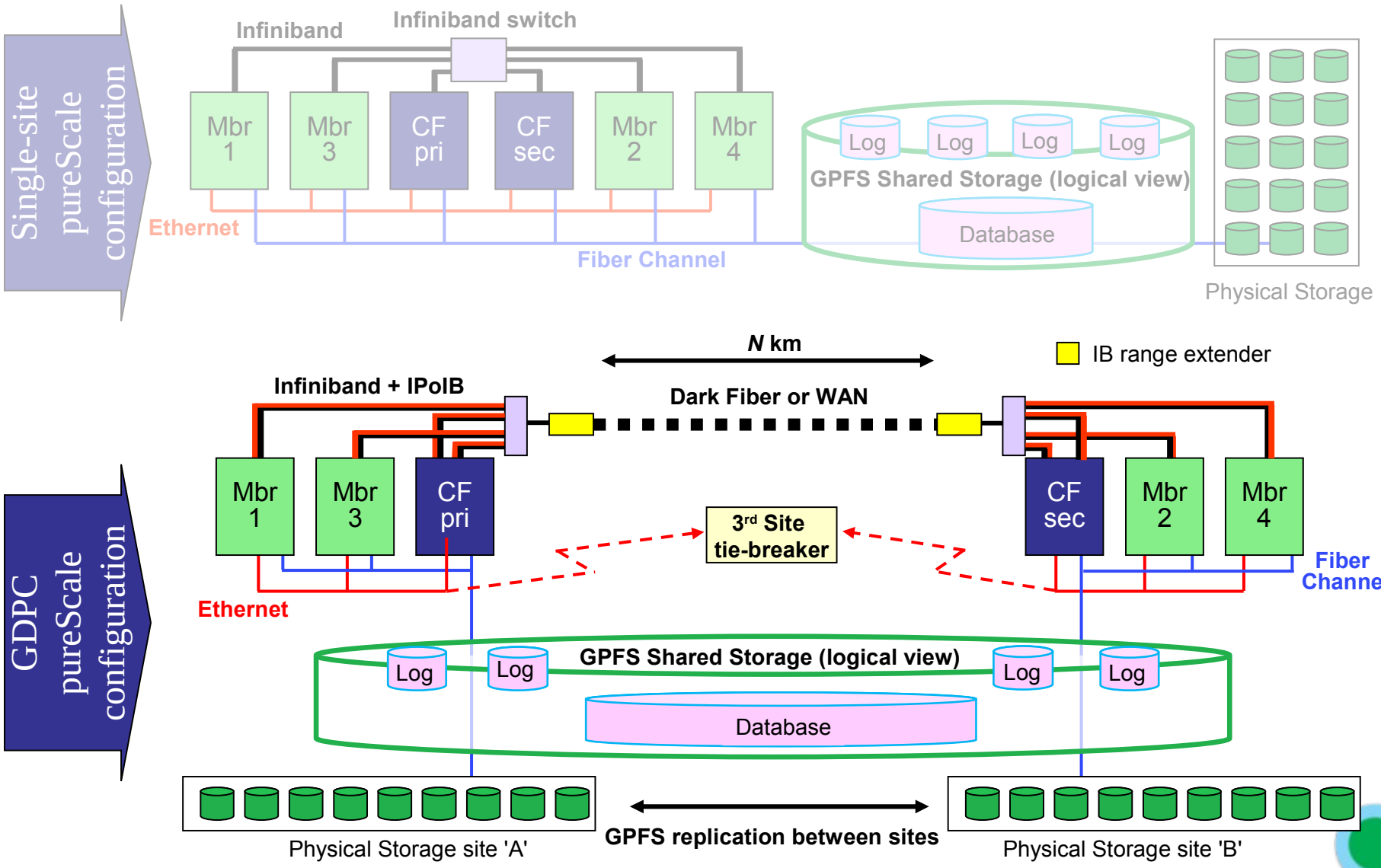
  http://www-03.ibm.com/systems/z/advantages/gdps/index.html

Target scenario

- Both sites are active & available for transactions during normal operation
- In the event of a failure, client connections are automatically redirected to surviving members by Workload Balancing (WLB) and Automatic Client Reroute (ACR)
  - Applies to both individual members within sites, and total site failure



Comparing single-site & GDPC pureScale configurations



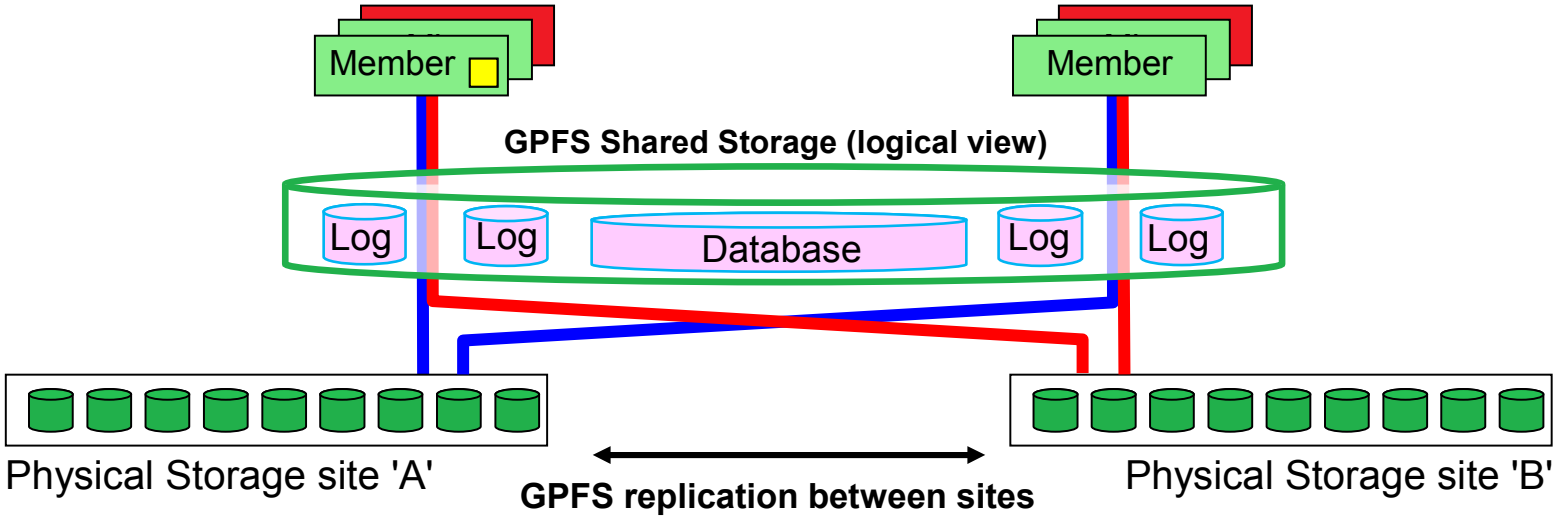
Long-distance Infiniband?

- Typical Infiniband connectivity reaches at most 10-20 m
  - Specialized cables allow up to a few hundred meters
- DB2/z GDPS achieves long distances with specialized HCA2-O LR optical coupling adapter + repeaters
- pureScale GDPC uses IBTA-compliant range extenders
- For example, Obsidian 'Longbow' extenders
  http://www.obsidianresearch.com/products/e-series.html
  - Used in pairs, appear in network as a 2-port IB switch
  - Convert duplex IB traffic to dark fiber or 10 GbE WAN traffic

Longbow C-103

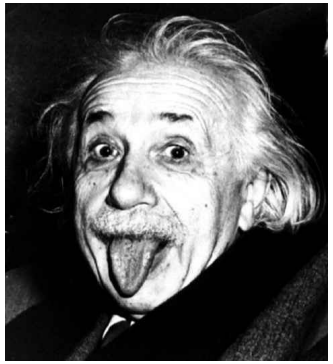
Longbow E-100

Disk storage in GDPC



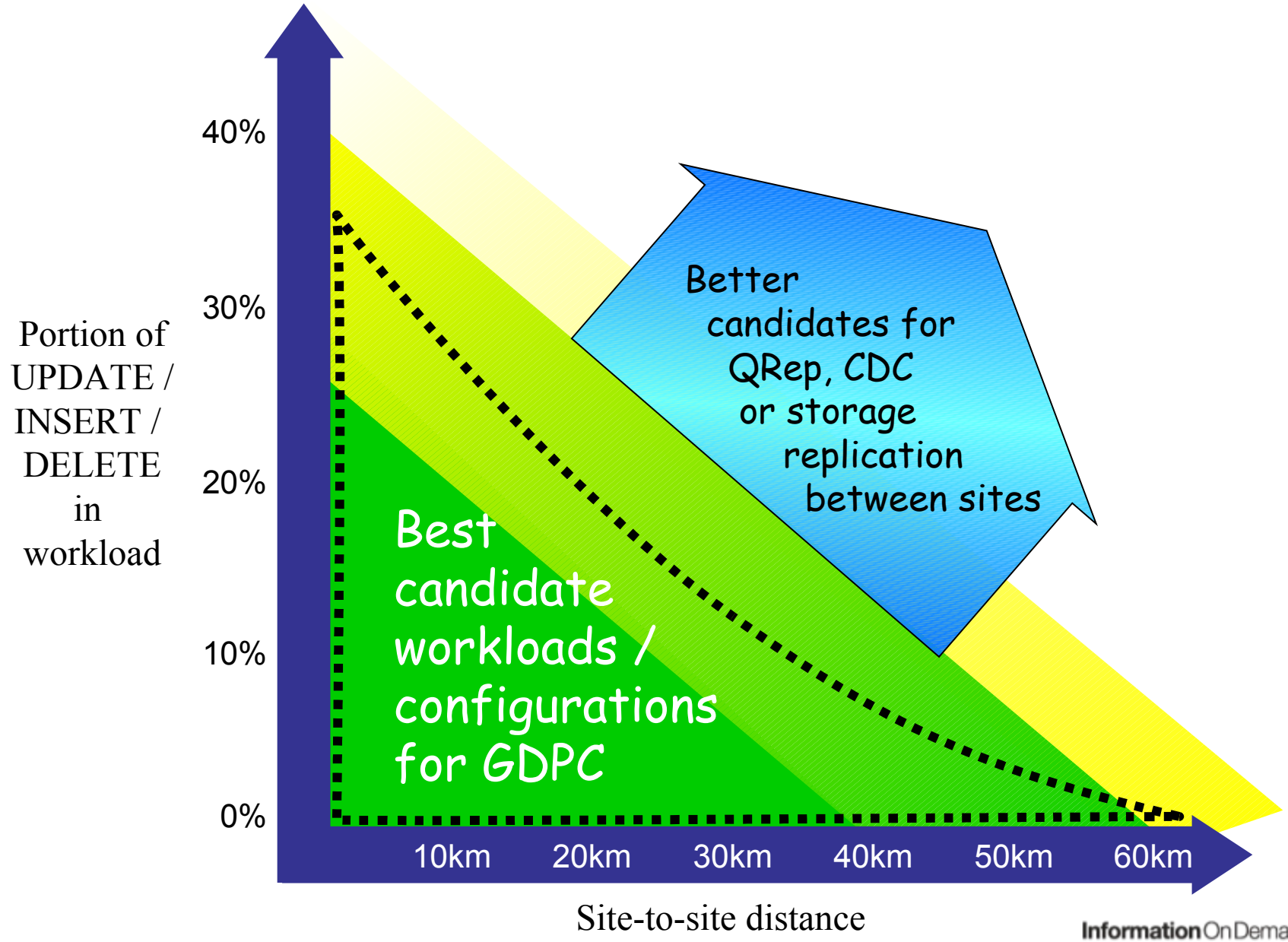
- GPFS replication coordinates synchronous writes across sites
  - Any write to the cluster storage from either site is replicated to the storage at the other site
- All storage is connected to pureScale hosts at both sites via zoned SANs
  - GPFS daemons on each server write to both site replicas directly – not by passing updated pages between GPFS daemons
- Replication of writes and site-to-site distance causes some increase in write times for both transaction logs and containers
- Reads are optimized to use the local copy for best performancep

Some characteristics of GDPC clusters



- Unavoidable increase in message latency
  - 5 µs / km limit in glass fiber due to speed of light
    - 30 µs round-trip from member to CF @ 3km
    - 100 µs round-trip at 10km, etc.
  - Greater if repeaters or "slow" WAN are used
  - Longer message latency can have a negative impact on cluster performance
- Workloads with a greater portion of read activity (SELECTs) vs. writes tend to see lower impact due to distance
  - GDPC is best suited for higher read content workloads (e.g. 80% or more read activity)
  - Impact of R/W ratio grows with distance between sites

The 'sweet spot' of GDPC



What do I need for a GDPC deployment?

1. Existing dark fiber (DWDM) or WAN connection between sites A & B
   - With required infrastructure (e.g. repeaters) for the distance involved
2. A third tie-breaker site with ethernet connectivity to sites A & B
   - Enables automatic recovery from complete failure of either site
3. One or two pairs of Infiniband extenders
   - Dual links / extender pairs can avoid single-point-of-failure and provide additional site-site capacity
4. SAN infrastructure to support GPFS replication between sites A & B
   - All storage must be 'visible' at both sites for access in the event of site failure
   - See GPFS redbook for additional details on GPFS replication
5. Client connectivity to sites A & B

For information on services required for deployment

- Contact go_db2@ca.ibm.com

More information about GDPC configurations

https://www.ibm.com/developerworks/data/library/long/dm-1104purescalegdpc/

- Concepts
- Comparisons with single-site pureScale cluster configurations
- Step-by-step setup instructions