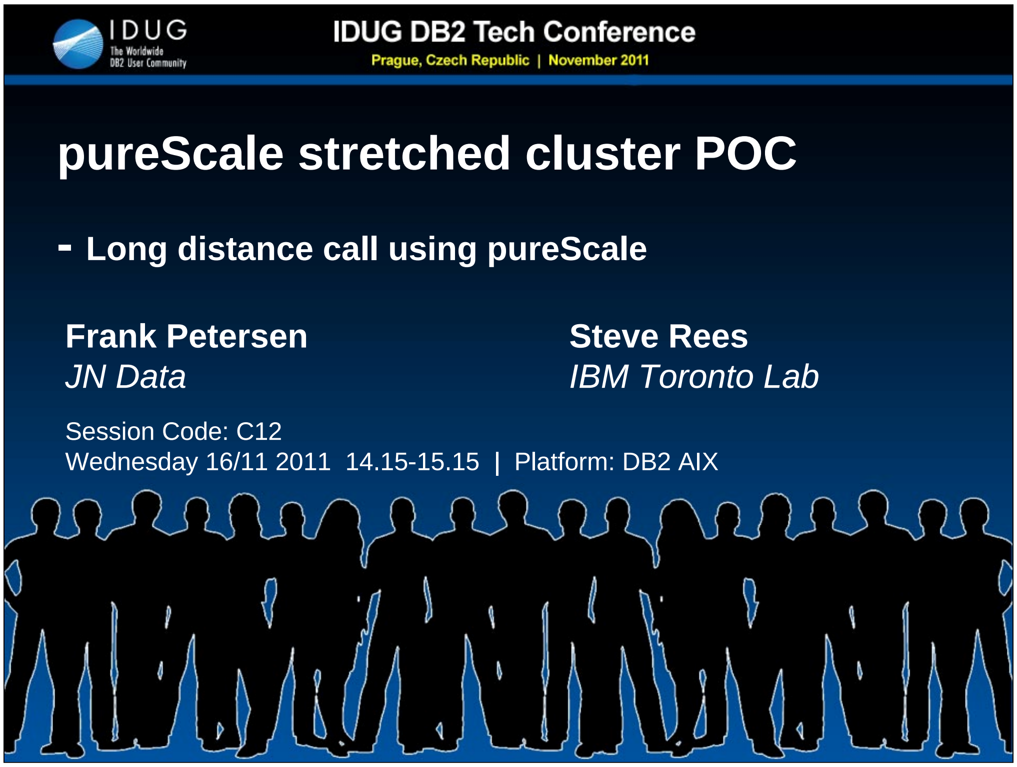


IDUG DB2 Tech Conference

Prague, Czech Republic | November 2011

pureScale stretched cluster POC

- Long distance call using pureScale

Frank Petersen
JN Data

Steve Rees
IBM Toronto Lab

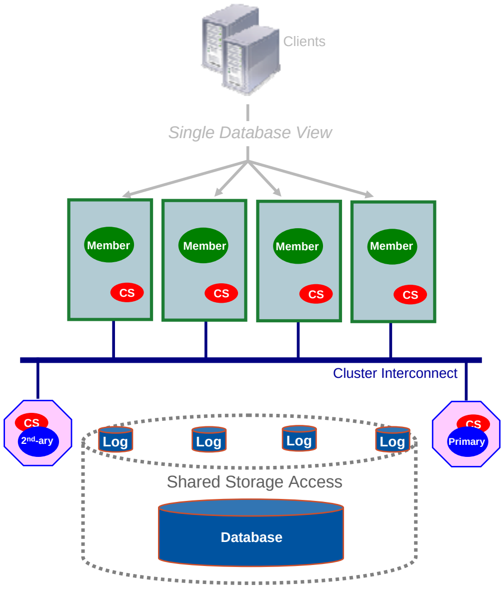
Session Code: C12
Wednesday 16/11 2011 14.15-15.15 | Platform: DB2 AIX

This presentation covers DB2 pureScale geographically dispersed clusters. pureScale is DB2 LUW’s answer to DataSharing in DB2 z/OS to provide DB2 LUW with unlimited scalability and high availability. Until now, pureScale clusters have only been supported in configurations where the physical boxes were placed within a very limited distance of one another. By exploiting some very advanced network technology, pureScale clusters can now be 'stretched' so that they can be used in an installation where 2 sites are placed kilometers apart. In this way pureScale can take part in an "active/active" disaster recovery (DR) setup, where one surviving site can take over the workload should the other site fail, and in this way ensure maximum availability. In 2011 Bankdata/JN Data and IBM started a “Proof of Concept” to bring this setup into the Bankdata/JN Data installation. This presentation will be covering both the physical (hardware) setup, the software setup and information about all the tests of the error scenarios that were performed at Bankdata/JN Data to verify the solution.

Agenda

- Part 1 - Introduction to DB2 pureScale 'stretch' cluster (GDPC)
- Part 2 - The Bankdata PoC
  - Who and why
  - The target application
  - Setup –
    - HW
    - prereq SW
    - pureScale
  - Test plan
  - Experiences
  - Conclusions

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

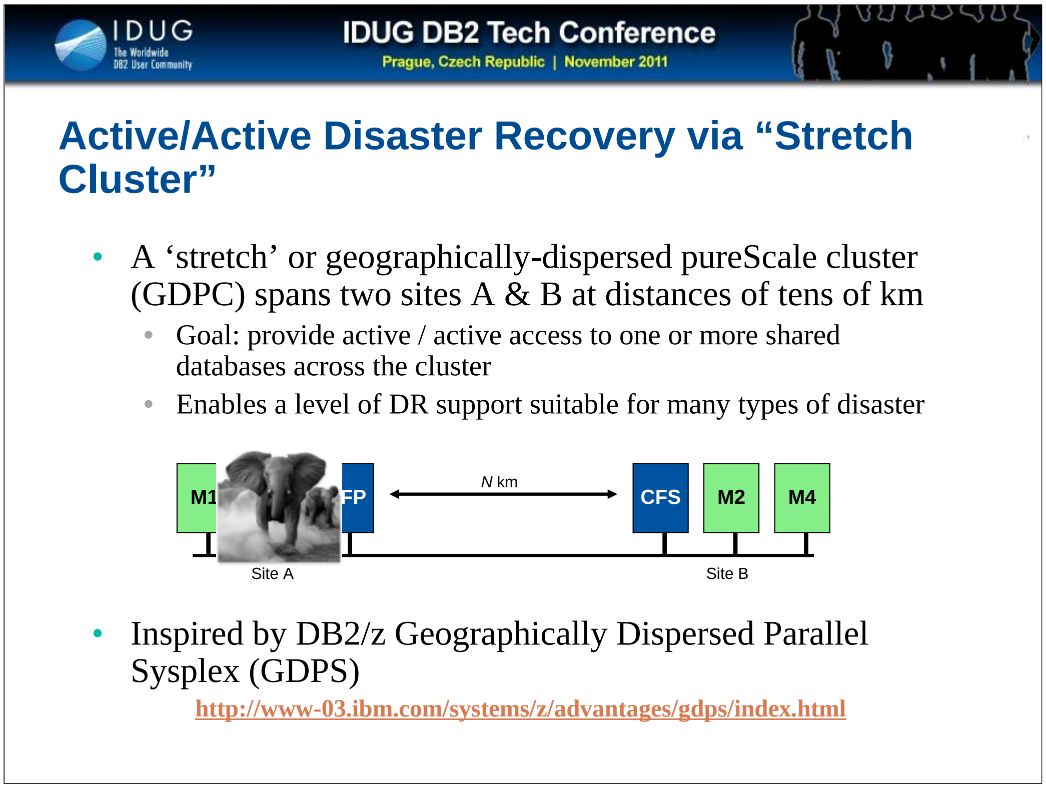
- Special optimizations provide significant advantages on RDMA-capable interconnects like Infiniband

Cluster Caching Facility (CF) technology

- 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)



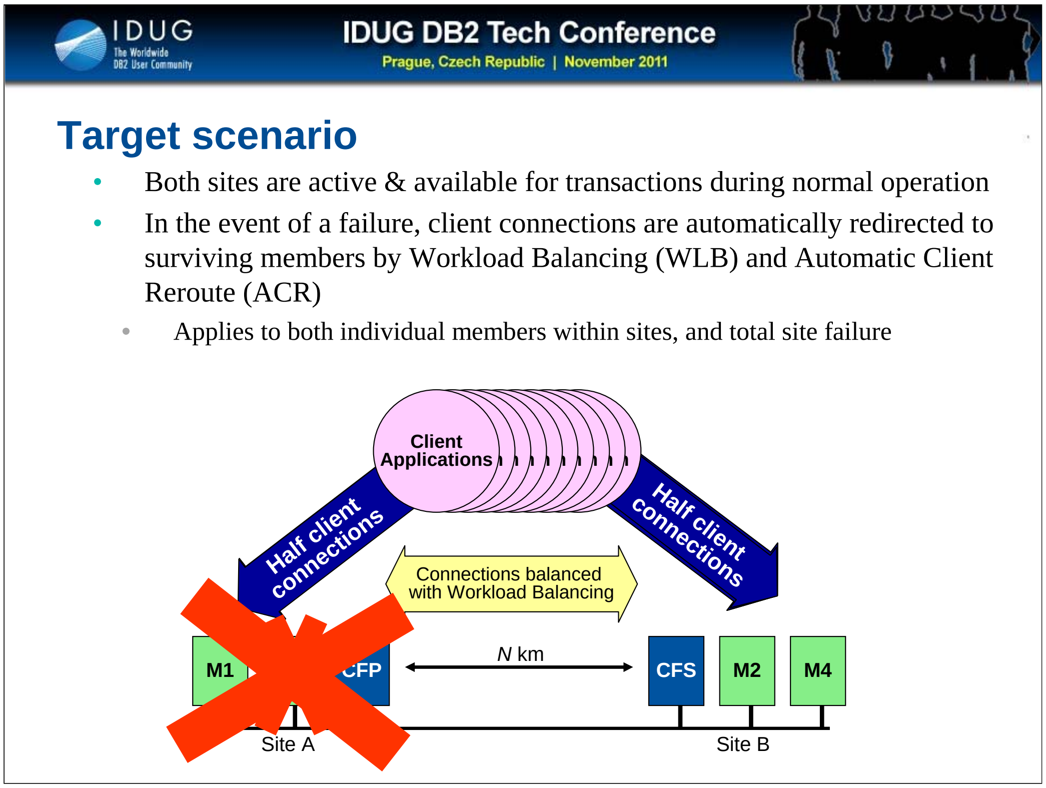
Active/Active Disaster Recovery via “Stretch Cluster”

- 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

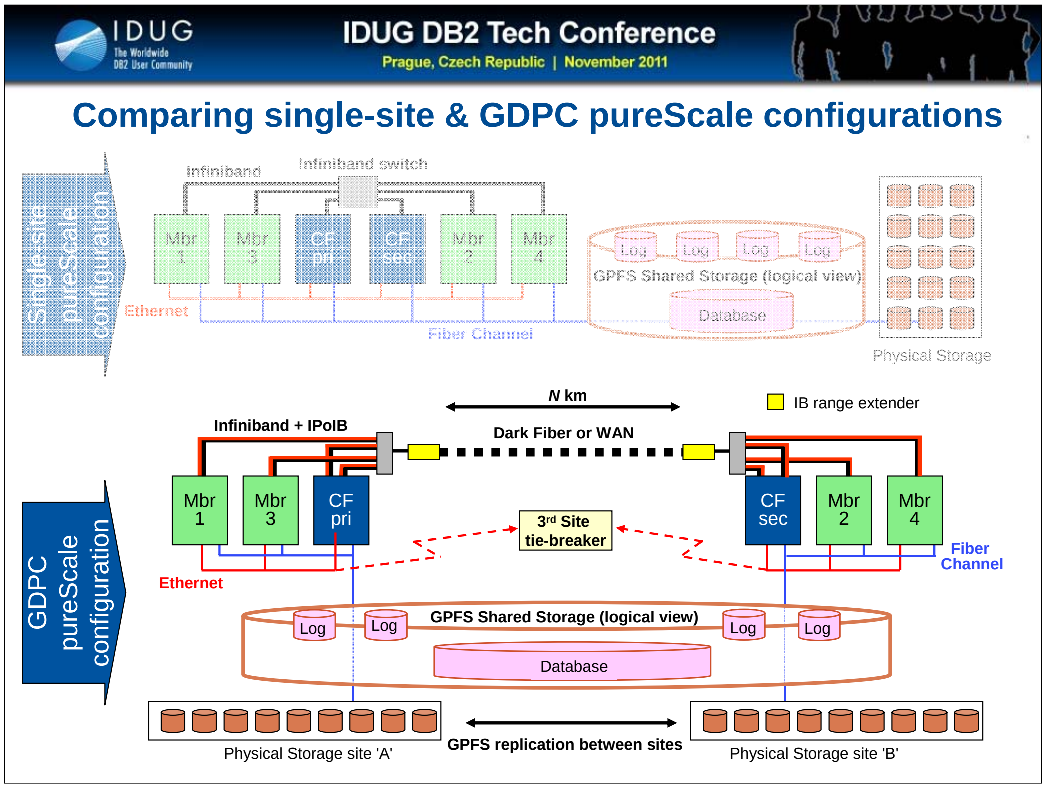
The 'geographically dispersed pureScale cluster' allows fully synchronized read/write activity concurrently at both sites. If one site should happen to go down (say, due to charging elephants, or other more prosaic problems), the other half of the cluster remains functional to continue work.



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

Different levels of failures – from single or multiple members, members and CF, or even an entire site – can be handled by the GDPC configuration.



IDUG The Worldwide DB2 User Community

IDUG DB2 Tech Conference

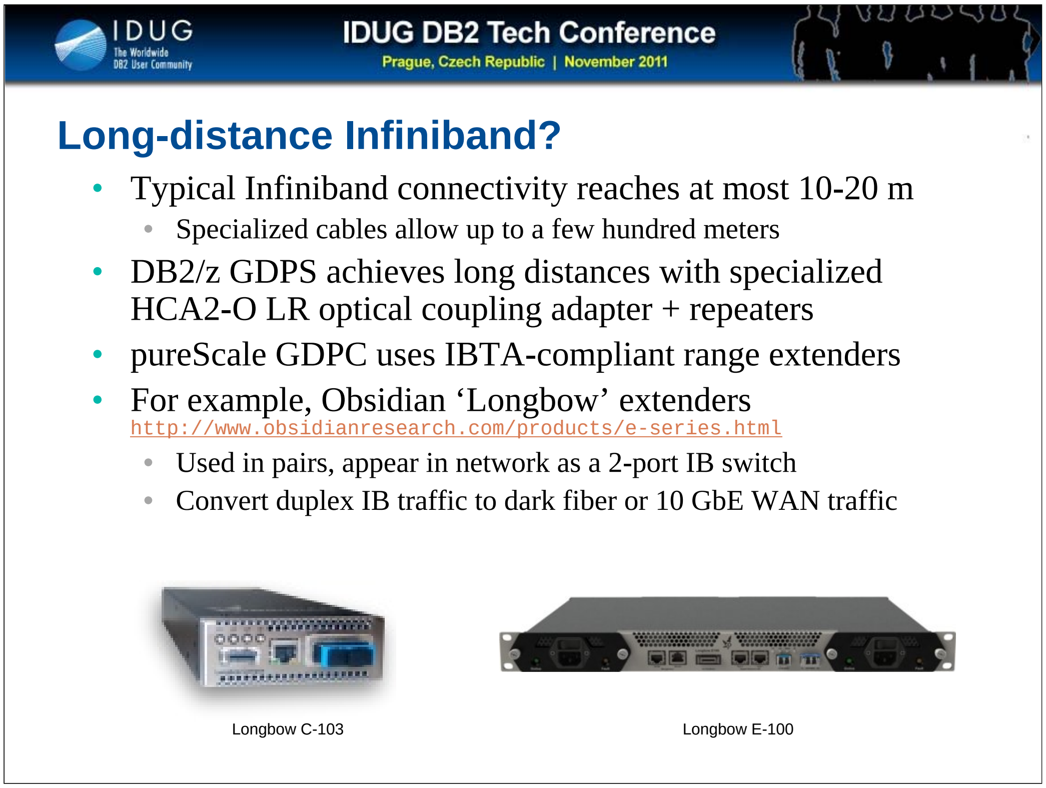
Prague, Czech Republic | November 2011

Comparing single-site & GDPC pureScale configurations

In simplest terms, GDPC basically splits a regular pureScale cluster, putting half of the compute resource at each site. Communication is maintained with advanced Infiniband extender technology, and synchronized versions of the on-disk data are maintained transparently using GPFS synchronous replication.

The 3rd site tie-breaker is required to avoid 'split brain' cases, where the network between sites might go down, and neither side can legitimately claim to be "THE" cluster afterward (or worse, they both do.) The tie-breaker is very modest – it does not have to have access to the SAN or IB. The tie-breaker *could* be located at either of the two main sites – but then there are challenges if both site A (for example) and the tiebreaker go down. Extra steps will be required to bring up the remaining site, since quorum can't be achieved with both A and tiebreaker down.

For details on GPFS synchronous replication, see GPFS Admin doc SC23-5182-02



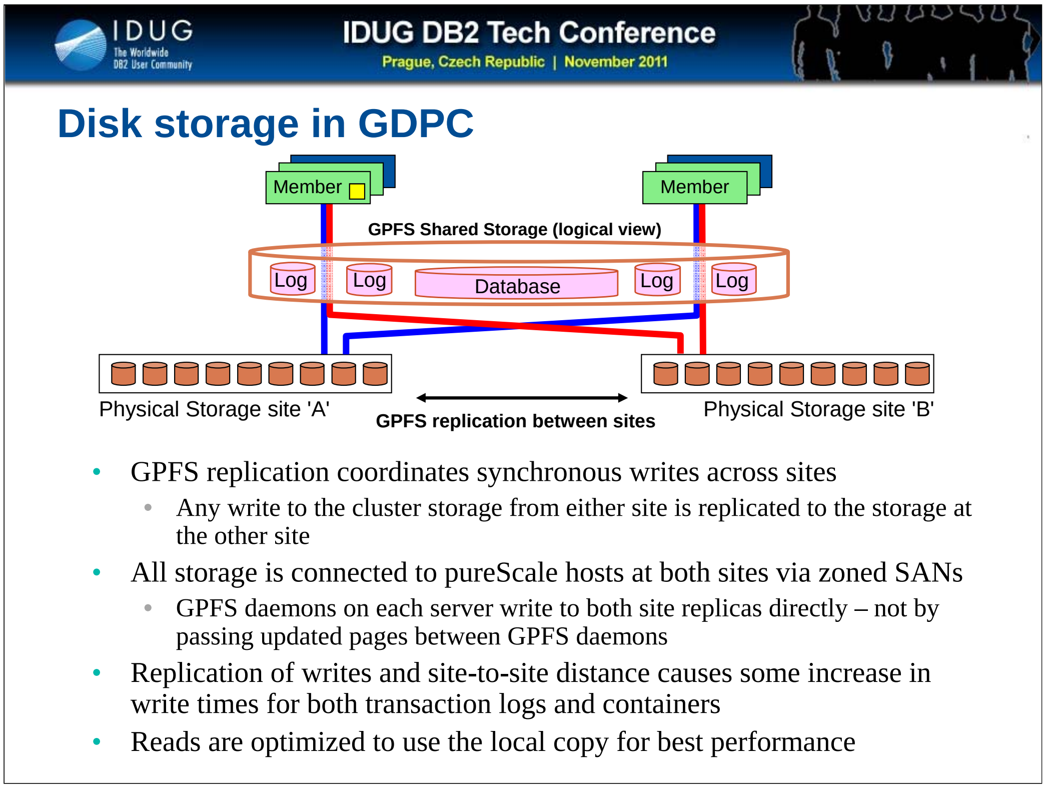
Long-distance Infiniband?

- 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

Long-distance Infiniband exists in the DB2 Sysplex world already, providing support for mainframe-based distributed clusters (GDPS). pureScale utilizes IBTA-standard Infiniband adapters and switches, which can be extended over long distances with devices such as the Obsidian Longbow IB extenders.



IDUG The Worldwide DB2 User Community

IDUG DB2 Tech Conference

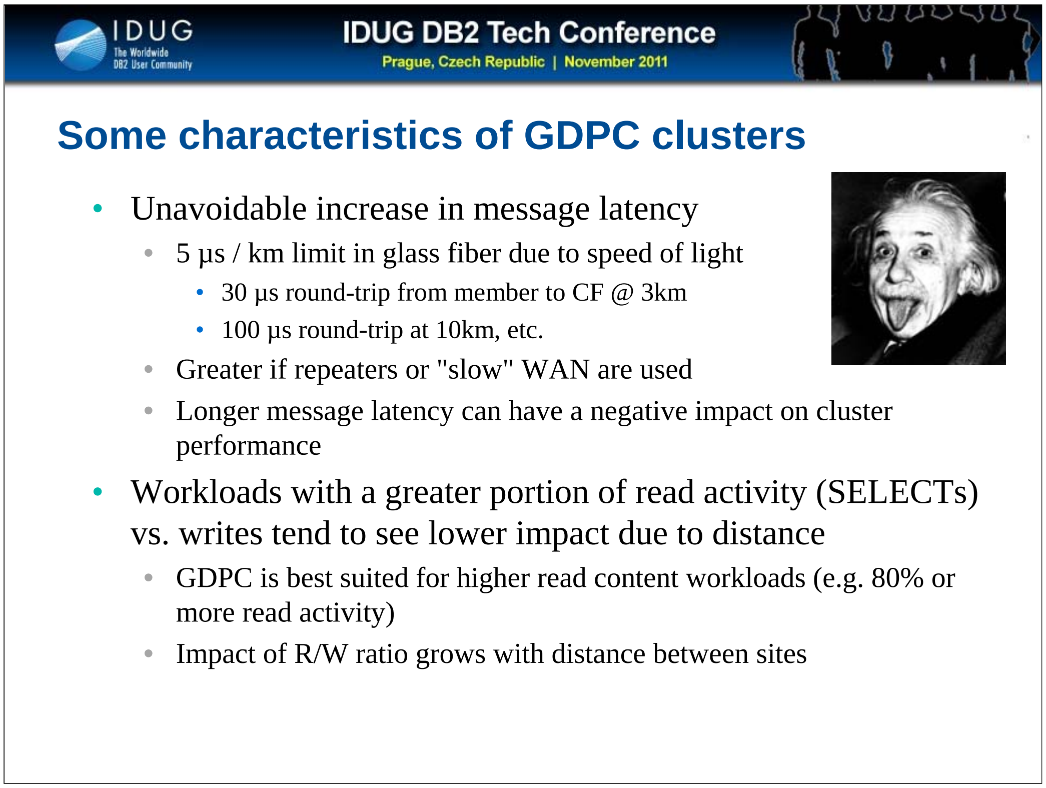
Prague, Czech Republic | November 2011

Disk storage in GDPC

Member
Member
GPFS Shared Storage (logical view)
Log
Log
Database
Log
Log
Physical Storage site 'A'
GPFS replication between sites
Physical Storage site 'B'

- 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 performance

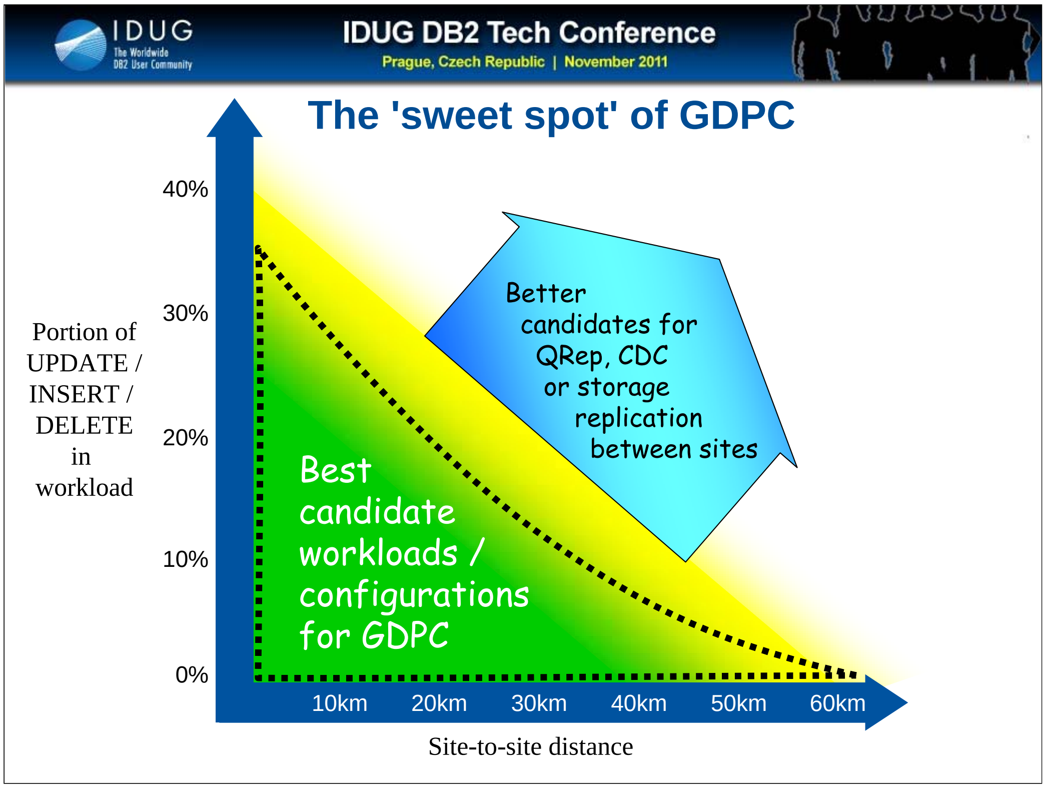
DR isn't DR unless each site can continue operation without the other. And that means that two copies of the database must exist – one at each site. GPFS synchronous replication keeps them in sync, but a zoned SAN setup provides the underlying infrastructure for the systems at each site to access the disks at both sites.



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

Much as we'd like it otherwise, the speed of light is finite. So once we start adding distance between sites, messaging delays due to the speed of light start to creep in. These can be quite insigificant at very short distances, however they can add up as the sites get to be 20 or 30 or more km apart. Because database write operations require more message traffic on average than read operations, the nature of the workload (read heavy, or write heavy?) is an additional factor in maximum practical distance between sites.



IDUG The Worldwide DB2 User Community

IDUG DB2 Tech Conference

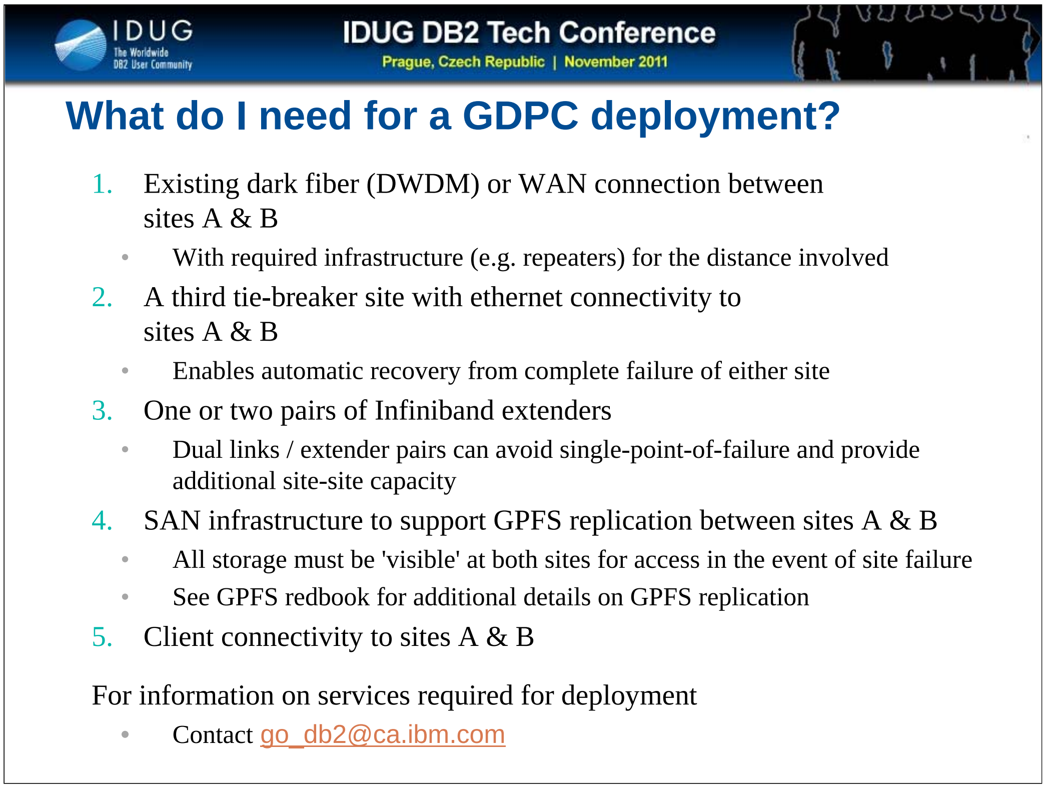
Prague, Czech Republic | November 2011

The 'sweet spot' of GDPC

Better candidates for QRep, CDC or storage replication between sites

Best candidate workloads / configurations for GDPC

pureScale supports GDPC as well as other DR solutions, such as Q Replication (QRep) and Change Data Capture (CDC). The 'sweet spot' of GDPC, where it's the most suitable choice, typically involves relatively close site-to-site distances, and higher read ratios.



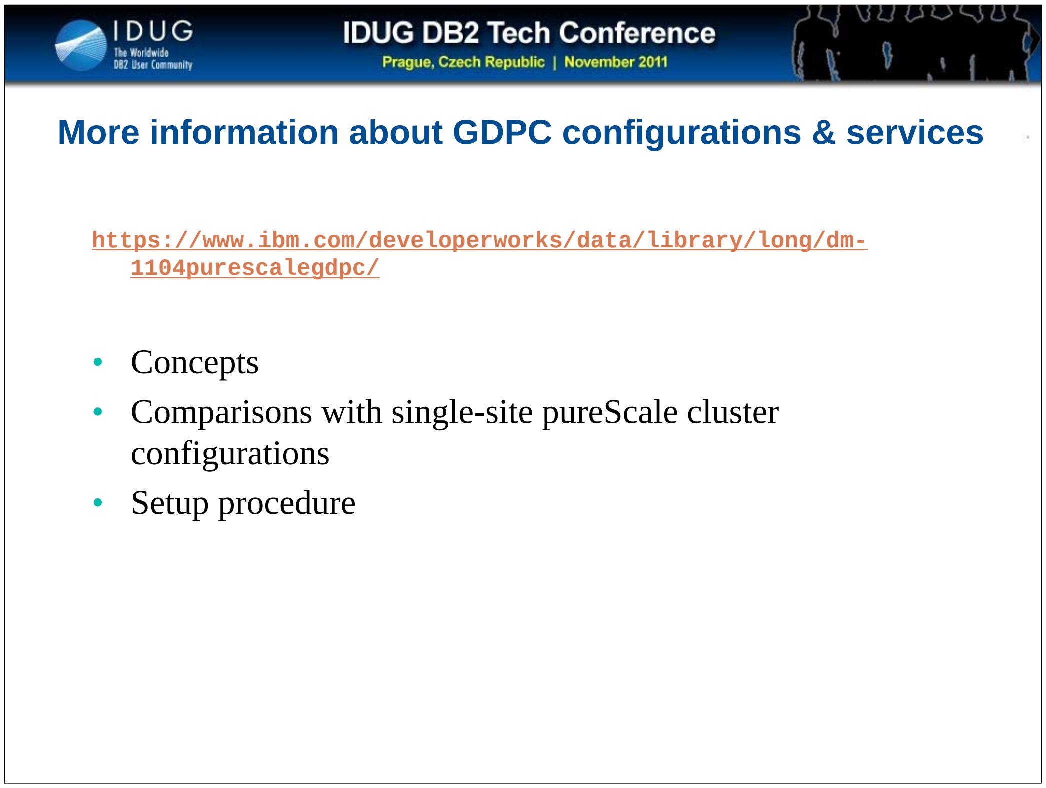
What do I need for a GDPC deployment?

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

GDPC is fundamentally a 'normal' pureScale cluster stretched over two sites – so the core system requirements of hardware and software are very similar. The main differences on top of this include (1) a high-bandwidth, low-latency WAN or dark fiber connection between sites, (2) Infiniband extenders to span that distance, (3) zoned SAN storage to provide disaster toleration between sites, and (4) GPFS replication to keep storage content in sync.



IDUG The Worldwide DB2 User Community

IDUG DB2 Tech Conference

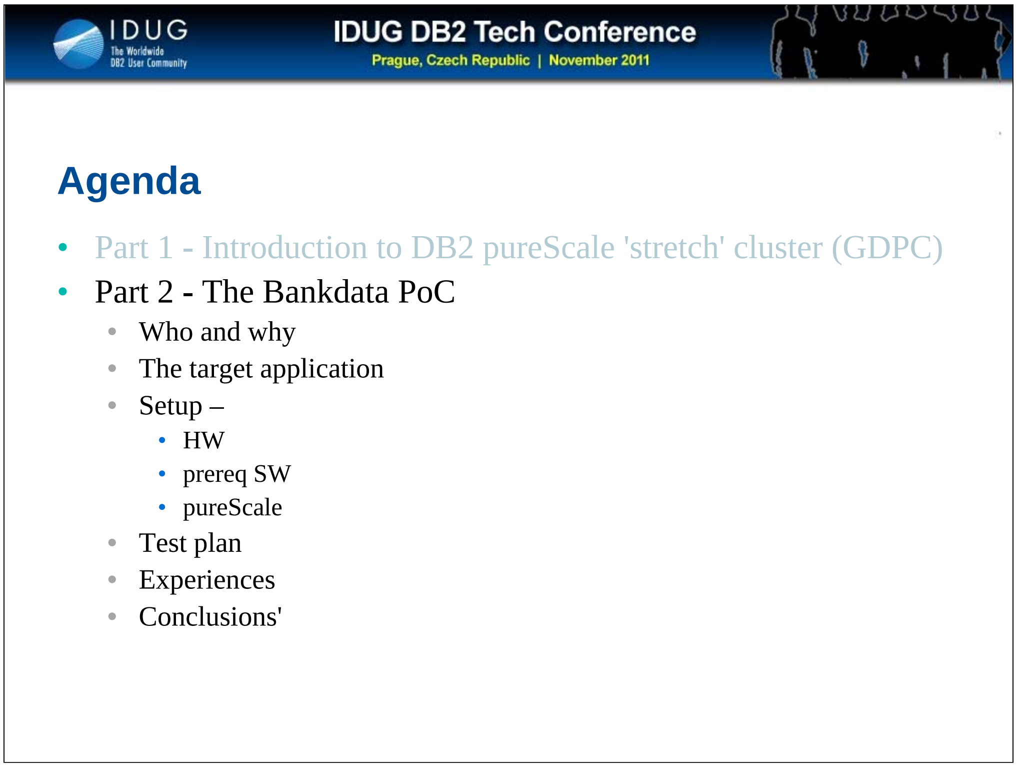
Prague, Czech Republic | November 2011

More information about GDPC configurations & services

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

- Concepts
- Comparisons with single-site pureScale cluster configurations
- Setup procedure

The referenced whitepaper goes into detail on how a GDPC is configured, how it's different from a regular pureScale cluster, etc. Definitely required reading for anyone interested in a GDPC deployment!



Agenda

- Part 1 - Introduction to DB2 pureScale 'stretch' cluster (GDPC)
- Part 2 - The Bankdata PoC
  - Who and why
  - The target application
  - Setup –
    - HW
    - prereq SW
    - pureScale
  - Test plan
  - Experiences
  - Conclusions'

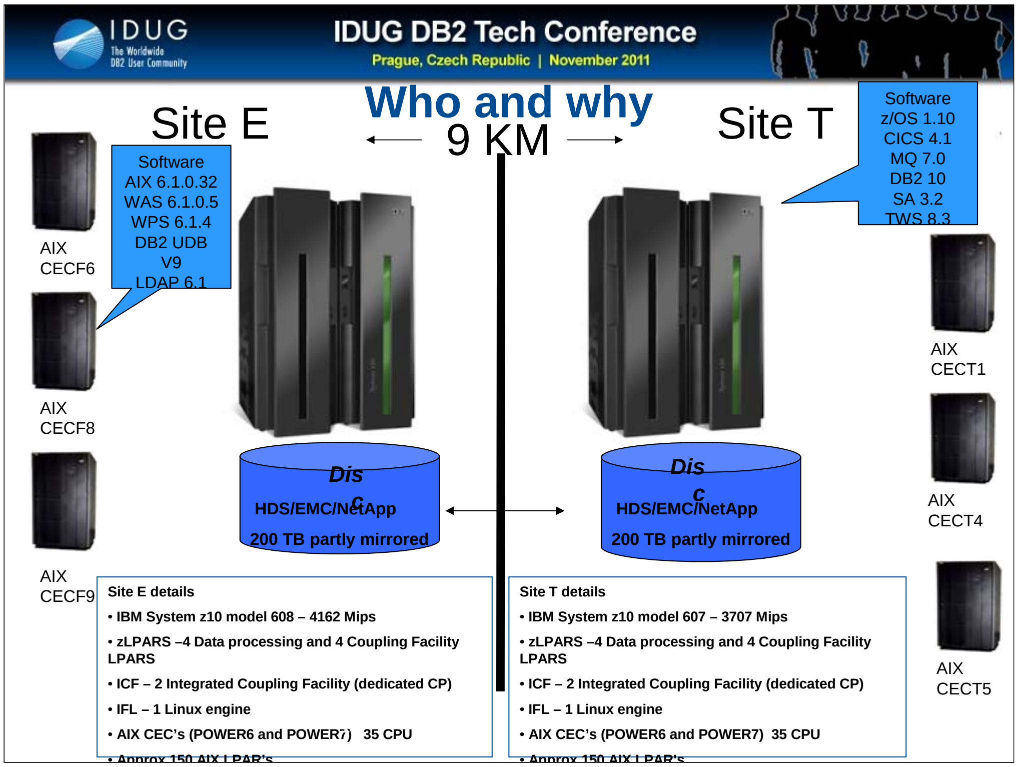
WHO ? - Location and history

- Bankdata is located in the city of Fredericia, Denmark
- Bankdata was founded on June 16th, 1966
- The first data processing action was made on October 18th, 1967
- Running IT for 14 Danish banks – number 15 on it's way…..
- Currently 500 employees

- Bankdata is organized as an association with the purpose of developing IT solutions, data processing and related activities for the members of the association and for service members

- Bankdata is owned 100% by the members of the association - our customers are our owners!

- Non-profit centre !!



IDUG The Worldwide DB2 User Community

IDUG DB2 Tech Conference

Prague, Czech Republic | November 2011

Who and why

Site E ← 9 KM → Site T

Software
AIX 6.1.0.32
WAS 6.1.0.5
WPS 6.1.4
DB2 UDB V9
LDAP 6.1

Software
z/OS 1.10
CICS 4.1
MQ 7.0
DB2 10
SA 3.2
TWS 8.3

Dis
c
HDS/EMC/NetApp
200 TB partly mirrored

Dis
c
HDS/EMC/NetApp
200 TB partly mirrored

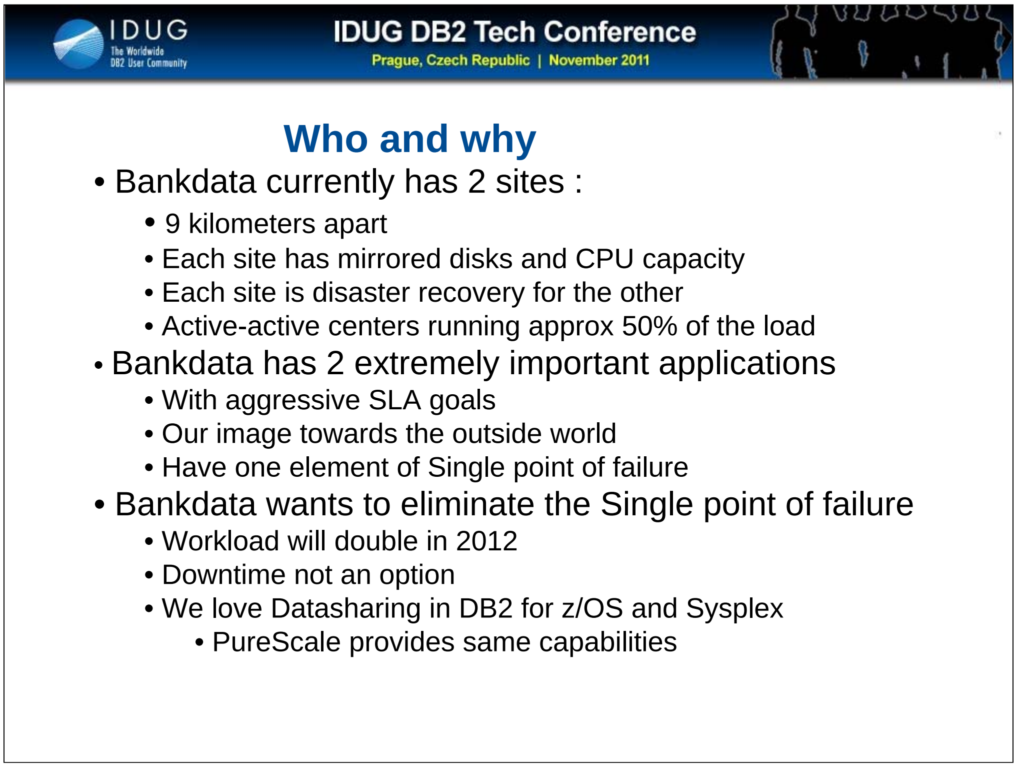
Site E details

- **IBM System z10 model 608 – 4162 Mips**
- **zLPARS –4 Data processing and 4 Coupling Facility LPARS**
- **ICF – 2 Integrated Coupling Facility (dedicated CP)**
- **IFL – 1 Linux engine**
- **AIX CEC's (POWER6 and POWER7) 35 CPU**
- **Approx 150 AIX LPAR's**

Site T details

- **IBM System z10 model 607 – 3707 Mips**
- **zLPARS –4 Data processing and 4 Coupling Facility LPARS**
- **ICF – 2 Integrated Coupling Facility (dedicated CP)**
- **IFL – 1 Linux engine**
- **AIX CEC's (POWER6 and POWER7) 35 CPU**
- **Approx 150 AIX LPAR's**

Hardware and computer center physical layout.



IDUG DB2 Tech Conference

Prague, Czech Republic | November 2011

Who and why

- Bankdata currently has 2 sites :
    - 9 kilometers apart
    - Each site has mirrored disks and CPU capacity
    - Each site is disaster recovery for the other
    - Active-active centers running approx 50% of the load
- Bankdata has 2 extremely important applications
    - With aggressive SLA goals
    - Our image towards the outside world
    - Have one element of Single point of failure
- Bankdata wants to eliminate the Single point of failure
    - Workload will double in 2012
    - Downtime not an option
    - We love Datasharing in DB2 for z/OS and Sysplex
        - PureScale provides same capabilities

Bankdata has an active-active politics concerning computer centers. This means that for a “level-1” disaster (if there is such a thing) where we “just” lose one center processing should be able to continue unaffected. On z/OS this is pretty easy as we have SYSPLEX and DB2 DataSharing etc so by placing sufficient mainframe capacity and mirroring all disks this is pretty straight forward.

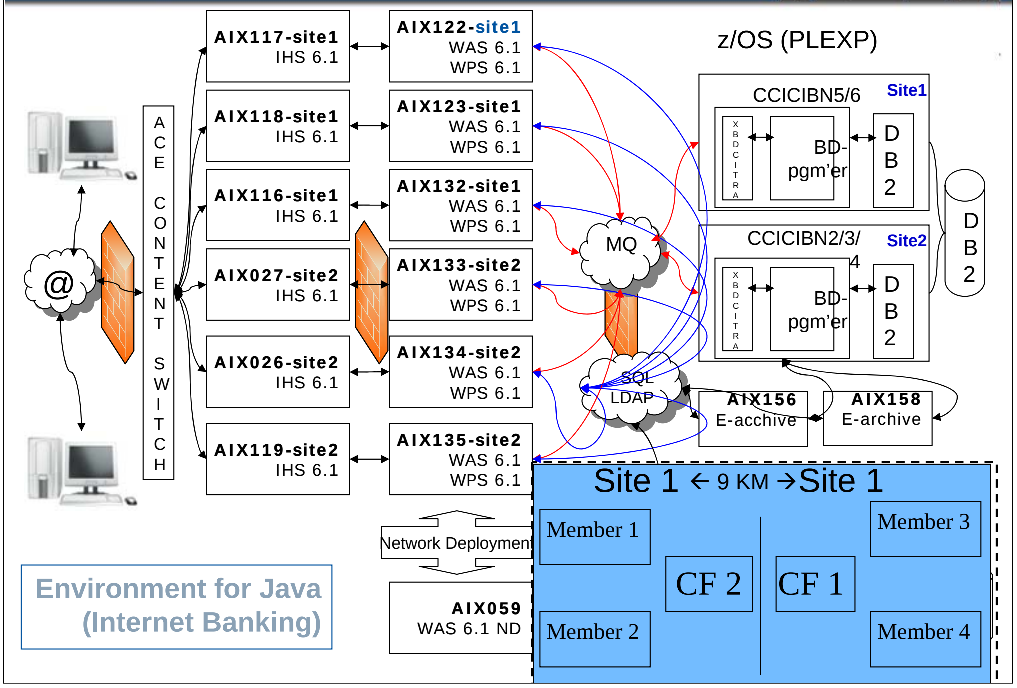
-

However on Windows and partly on AIX this is more difficult. Our main customer applications is running with most of the data on DB2 on z/OS using CICS transactions as “”WEB service calls””. Presentation layer and transaction driver is WAS/WPS running Java code on AIX. We have 50% of the WAS/WPS instances on either computer center but there is a small but highly active part of the data in DB2 UDB on AIX. Until now this has been a SPoF as we have had to make this an active/inactive DB2 UDB solution (“HACMP-like”). So if we loose one center we will have to wait for the DB2 UDB to be activated on the other side.

Similar upgrades and maintenance on the DB2 AIX will give a outage, though planned. We have in many years looked at DataSharing in DB2 on z/OS with envy when we are wearing our DB2 UDB glasses !!!

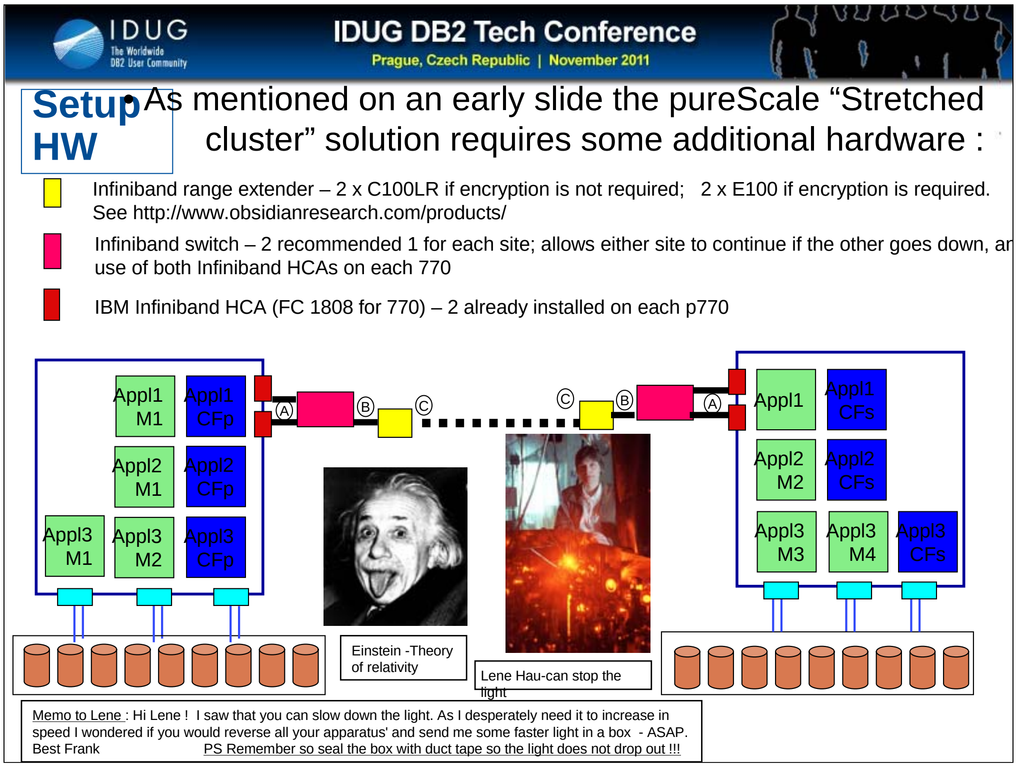
IDUG The Worldwide DB2 User Community

IDUG DB2 Tech Conference
Prague, Czech Republic | November 2011



Environment for Java (Internet Banking)

Just a picture on one of the main appliaction to highlight where the problem with SPoF is !!



IDUG The Worldwide DB2 User Community

IDUG DB2 Tech Conference

Prague, Czech Republic | November 2011

Setup HW

- As mentioned on an early slide the pureScale “Stretched cluster” solution requires some additional hardware :

🟨 Infiniband range extender – 2 x C100LR if encryption is not required;   2 x E100 if encryption is required. See http://www.obsidianresearch.com/products/

🟥 Infiniband switch – 2 recommended 1 for each site; allows either site to continue if the other goes down, ar use of both Infiniband HCAs on each 770

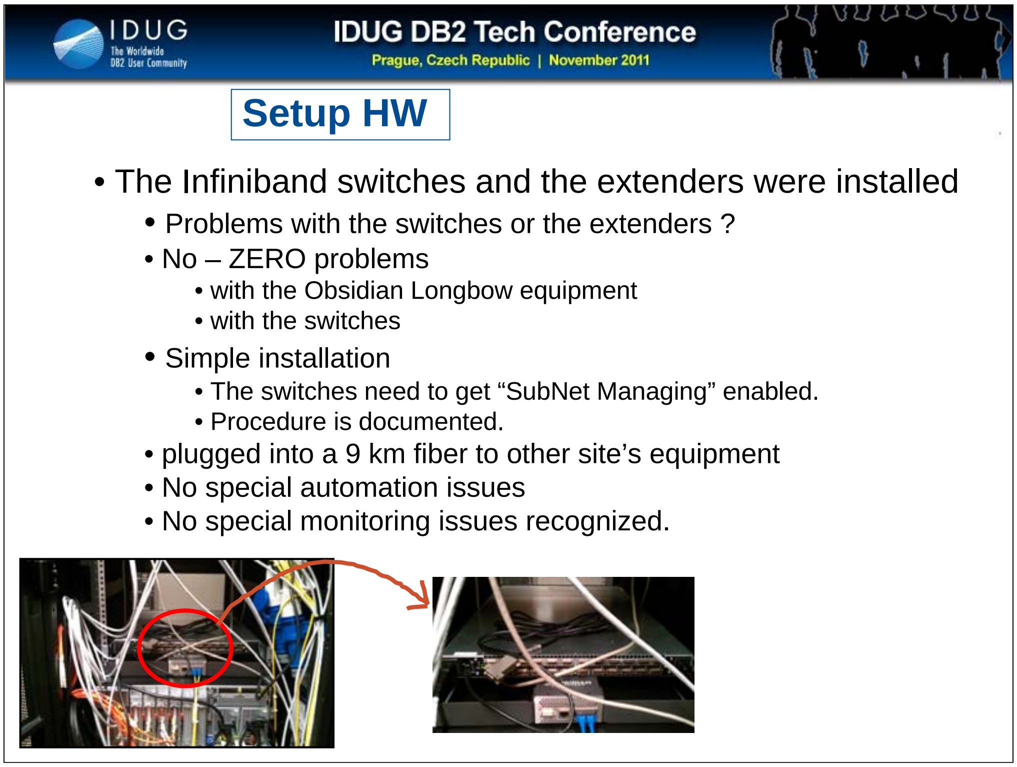
🟥 IBM Infiniband HCA (FC 1808 for 770) – 2 already installed on each p770

Einstein -Theory of relativity

Lene Hau-can stop the light

Memo to Lene : Hi Lene ! I saw that you can slow down the light. As I desperately need it to increase in speed I wondered if you would reverse all your apparatus' and send me some faster light in a box - ASAP.
Best Frank PS Remember so seal the box with duct tape so the light does not drop out !!!

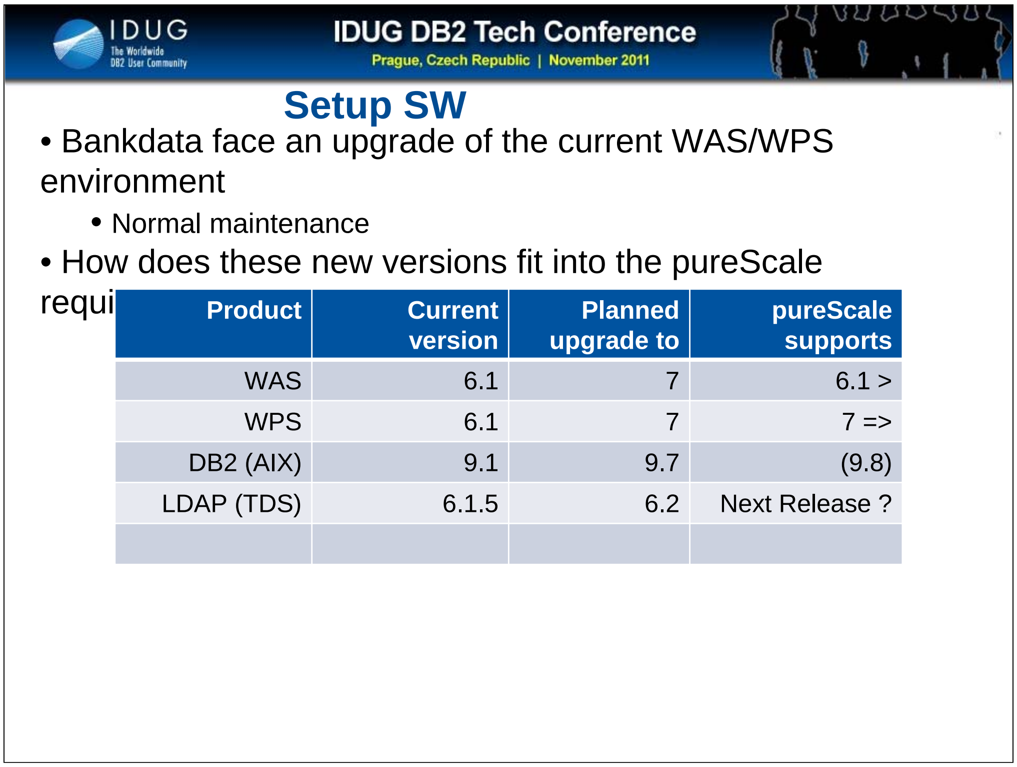
As understood by most people now pureScale brings DB2 on z/OS data sharing into DB2 UDB. Not implemented in the hardware but in a approximate version simulated in software components. Until now the two members of the cluster had to be ‘within a data center’ but with the Stretched Cluster solution WITH the usage of additional hardware we can now extent the distance – penalty being the speed of light …..



Setup HW

- The Infiniband switches and the extenders were installed
  - Problems with the switches or the extenders ?
  - No – ZERO problems
    - with the Obsidian Longbow equipment
    - with the switches
  - Simple installation
    - The switches need to get “SubNet Managing” enabled.
    - Procedure is documented.
  - plugged into a 9 km fiber to other site’s equipment
  - No special automation issues
  - No special monitoring issues recognized.

The Obsidian Longbow equipment is beautiful ! I am fond of beautifully solutions and I felt quite sad when we plugged all the fibers into them and stacked them in these messy racks ! Throughout the POC we observed no problems with all the equipment. We did no special monitoring because of this but we think that it is in fact possible to extract figures from the switches.



Setup SW

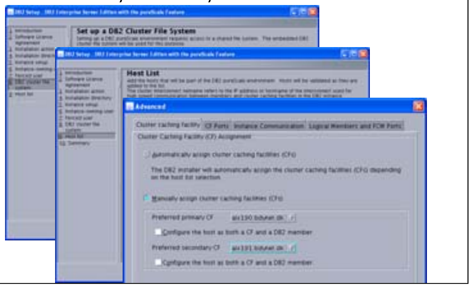
- Bankdata face an upgrade of the current WAS/WPS environment
  - Normal maintenance
- How does these new versions fit into the pureScale requi

Product	Current version	Planned upgrade to	pureScale supports
WAS	6.1	7	6.1 >
WPS	6.1	7	7 =>
DB2 (AIX)	9.1	9.7	(9.8)
LDAP (TDS)	6.1.5	6.2	Next Release ?

There are some requirement for the software supported by pureScale. In our environment we needed to upgrade WAS/WPS anyway so it did not present a real problem. The closest we came to a problem was TDS or LDAP where we could not get the required version. There are however several ways to get passed this……

Upgrade to pureScale

- When going to pureScale there are a few prereq's :
  - Installing the Infiniband adapters and switches
  - Installing the Obsidian extenders
  - Be ready to use automatic storage table spaces (AMS)
  - Be ready to use GPFS (knowledge and experience)
    - Make a node ready to become a tie-breaker node
  - Installing DB2 9.7 on a node where pureScale is going to reside
    - Might not be needed if going for the db2look/db2move approach
- The install is pretty straight forward :
  - Define 3 file systems on the GPFS : Instance, LOGs, Database
  - Follow the instructions carefully
    - db2setup+line commands
  - Understand that some files will be shared across the cluster
  - When finished you will have :
    - x members
    - 2 CF's
    - and running in a cluster.



DB2 Setup - DB2 Enterprise Server Edition with the pureScale Feature

Set up a DB2 Cluster File System

Host List

Advanced

Cluster Caching Facility (CF) Assignment

○ Automatically assign cluster caching facilities (CFs)

The DB2 installer will automatically assign the cluster caching facilities (CFs) depending on the host list selection.

● Manually assign cluster caching facilities (CFs)

Preferred primary CF: aix190

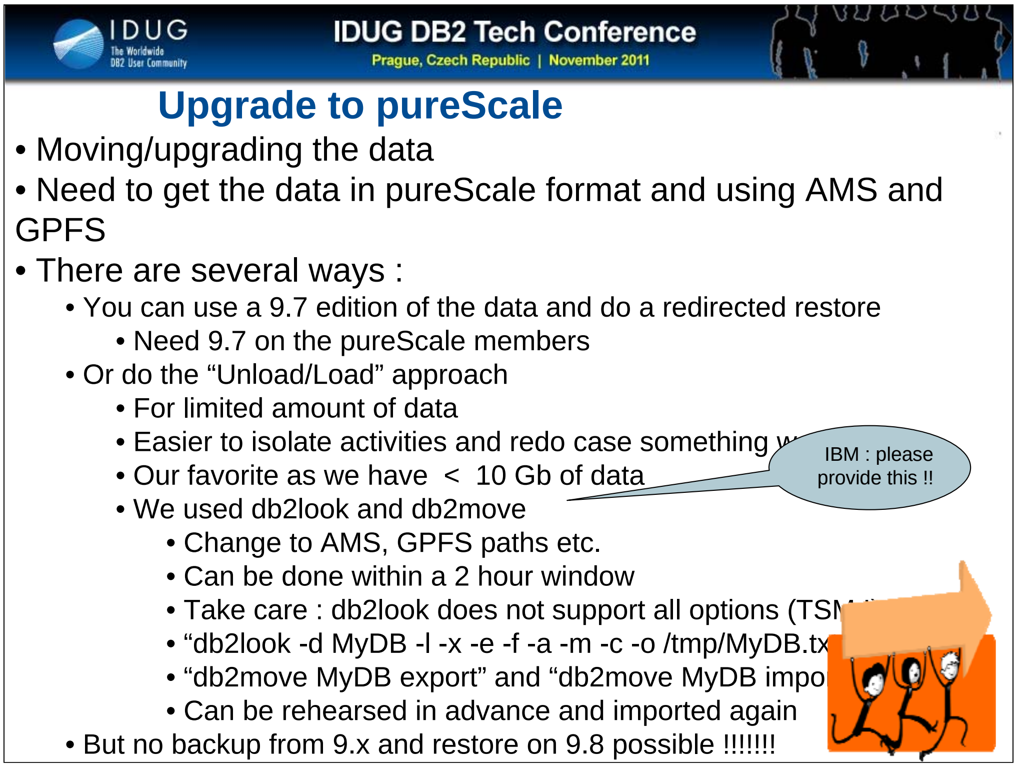
□ Configure the host as both a CF and a DB2 member

Preferred secondary CF: aix191

□ Configure the host as both a CF and a DB2 member

You have to realize that if you are a DB2 for z/OS datasharing freak things are a bit more complicated as the solution builds upon and around other building bricks and technology that you have to consider to which extent you have to master these….

For instance GPFS and if you have never user Automated Storage Managed Table Spaces. The thing is that the setup of much of this is done once and build as much into the pureScale installation procedure as possible but you still have to understand that in a disaster situation you have to be able to operate all these technology components.



IDUG The Worldwide DB2 User Community

IDUG DB2 Tech Conference

Prague, Czech Republic | November 2011

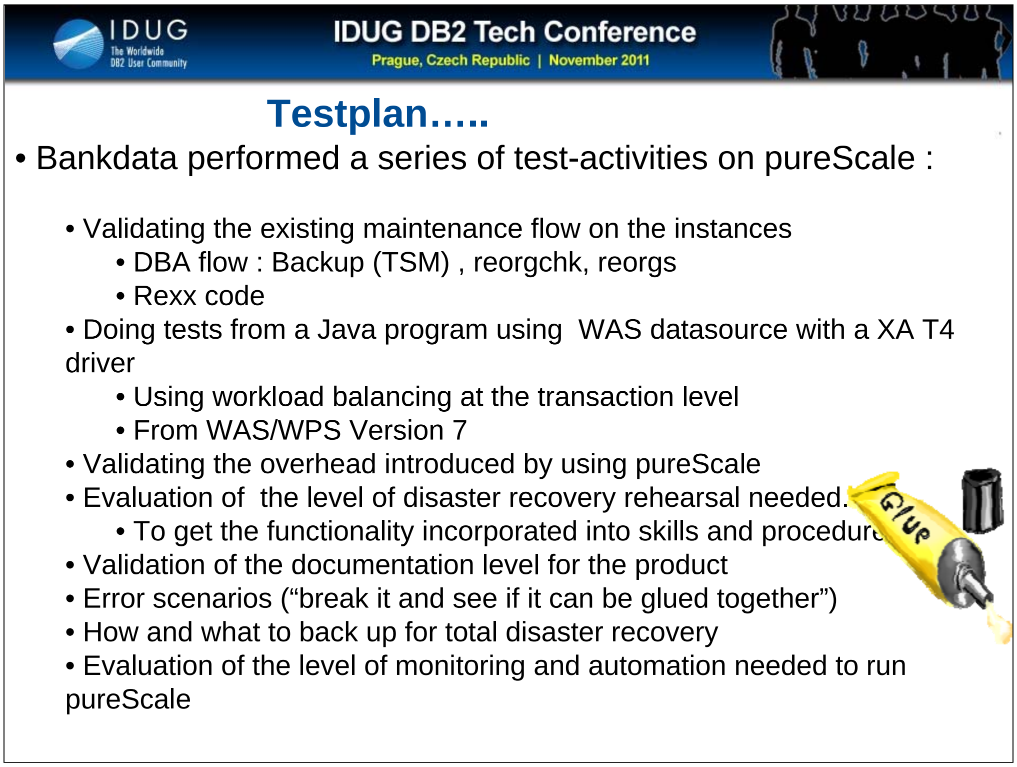
Upgrade to pureScale

- Moving/upgrading the data
- Need to get the data in pureScale format and using AMS and GPFS
- There are several ways :
  - You can use a 9.7 edition of the data and do a redirected restore
    - Need 9.7 on the pureScale members
  - Or do the “Unload/Load” approach
    - For limited amount of data
    - Easier to isolate activities and redo case something w
    - Our favorite as we have < 10 Gb of data
    - We used db2look and db2move
      - Change to AMS, GPFS paths etc.
      - Can be done within a 2 hour window
      - Take care : db2look does not support all options (TSM
      - “db2look -d MyDB -l -x -e -f -a -m -c -o /tmp/MyDB.tx
      - “db2move MyDB export” and “db2move MyDB impo
      - Can be rehearsed in advance and imported again
  - But no backup from 9.x and restore on 9.8 possible !!!!!!!

Now we need some data in our newly defined pureScale instance. The filesystems has to be GPFS and the table spaces AMS. If you will use the redirected restore you will need a DB2 9.7 on the system. Another way is to unload the data from the 9.x database and use a “load” on the pureScale instance. There are some clear advantages in this, for instance can you build the pureScale environment way before and do several test-loads before the final day arrives….

To do the unload/load a db2look and db2move is very elegant. You will however need to do some manual definitions at the pureScale side and understand that not all will be moved by db2look….

But it can be tested and rehearsed and build into a script. It would have been elegant if it was build into the pureScale installation as a script so every customer did not have to find the same pitfalls…. But it will soon be Christmas…



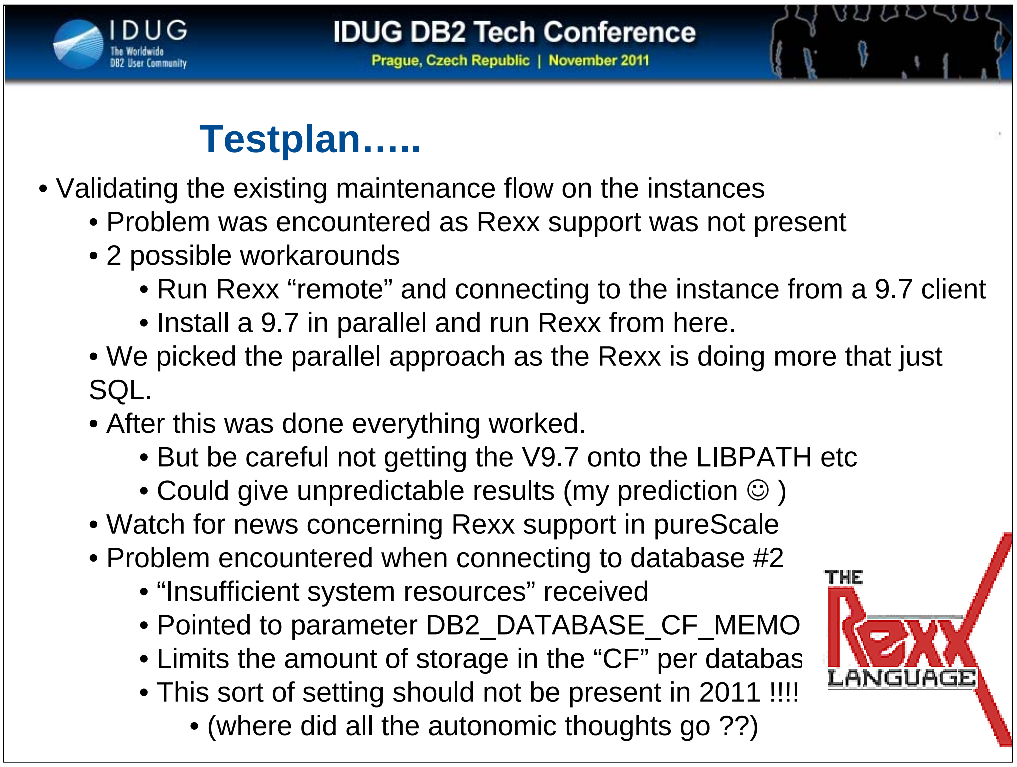
Testplan…..

- Bankdata performed a series of test-activities on pureScale :
  - Validating the existing maintenance flow on the instances
    - DBA flow : Backup (TSM) , reorgchk, reorgs
    - Rexx code
  - Doing tests from a Java program using WAS datasource with a XA T4 driver
    - Using workload balancing at the transaction level
    - From WAS/WPS Version 7
  - Validating the overhead introduced by using pureScale
  - Evaluation of the level of disaster recovery rehearsal needed.
    - To get the functionality incorporated into skills and procedure
  - Validation of the documentation level for the product
  - Error scenarios (“break it and see if it can be glued together”)
  - How and what to back up for total disaster recovery
  - Evaluation of the level of monitoring and automation needed to run pureScale

Okay, we have installed all the components and moved the data. Time is now right for testing the real reason why we do this : will pureScale let the workload continue with acceptable slowdown or interrupt without any manual intervention ? Will we see problems in the existing ‘housekeeping’ flow ? Will we impose an overhead on every transaction ? Is the doc understandable and sufficient ? Do we need additional monitoring ?

--

And most important will the systen survive all these more or less expected scenarios where we today will have an outage ??



IDUG The Worldwide DB2 User Community

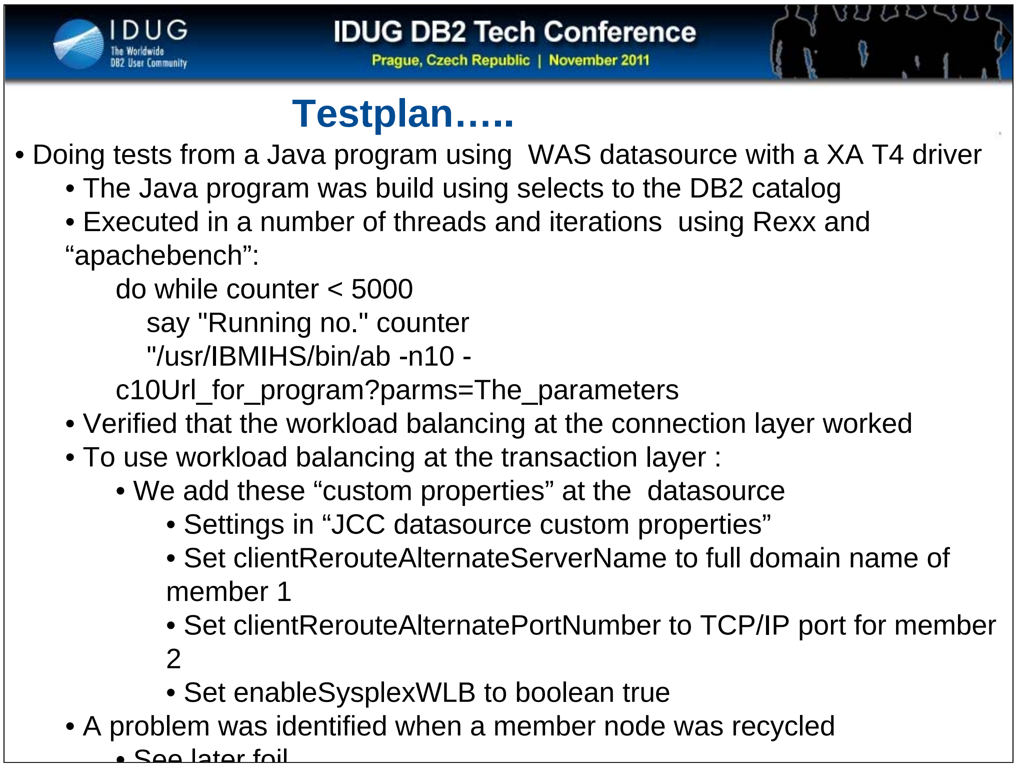
IDUG DB2 Tech Conference

Prague, Czech Republic | November 2011

Testplan…..

- Validating the existing maintenance flow on the instances
    - Problem was encountered as Rexx support was not present
    - 2 possible workarounds
        - Run Rexx “remote” and connecting to the instance from a 9.7 client
        - Install a 9.7 in parallel and run Rexx from here.
    - We picked the parallel approach as the Rexx is doing more that just SQL.
    - After this was done everything worked.
        - But be careful not getting the V9.7 onto the LIBPATH etc
        - Could give unpredictable results (my prediction ☺)
    - Watch for news concerning Rexx support in pureScale
    - Problem encountered when connecting to database #2
        - “Insufficient system resources” received
        - Pointed to parameter DB2_DATABASE_CF_MEMO
        - Limits the amount of storage in the “CF” per databas
        - This sort of setting should not be present in 2011 !!!!
            - (where did all the autonomic thoughts go ??)

Rexx code gave us some problems as this was not included in the version of pureScale that we tested. We expect it to be included in the near future but until them we will need a filesystem with the V9.7 code to allow Rexx. But do take care not to get V9.7 code into play in the pureScale instance !!!



Testplan…..

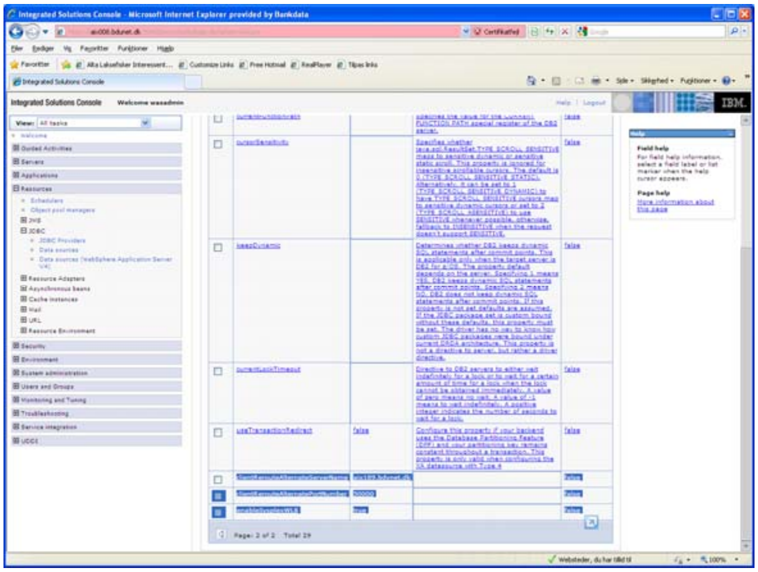
- Doing tests from a Java program using WAS datasource with a XA T4 driver
  - The Java program was build using selects to the DB2 catalog
  - Executed in a number of threads and iterations using Rexx and "apachebench":

```
do while counter < 5000
   say "Running no." counter
   "/usr/IBMIHS/bin/ab -n10 -
c10Url_for_program?parms=The_parameters
```

  - Verified that the workload balancing at the connection layer worked
  - To use workload balancing at the transaction layer :
    - We add these "custom properties" at the datasource
      - Settings in "JCC datasource custom properties"
      - Set clientRerouteAlternateServerName to full domain name of member 1
      - Set clientRerouteAlternatePortNumber to TCP/IP port for member 2
      - Set enableSysplexWLB to boolean true
  - A problem was identified when a member node was recycled
    - See later foil

When conducting test cases where you have to compare performance and verify specific scenarios you have 2 different routes to go. You can execute a 'real' workload or you can write some test programs that will fire of the canned set of workload. The later one will normally be the best way to go as it will allow you to redo the same scenario as often as you will to validate different settings and fix-levels. We made a small test driver simply doing some catalog lookup and executes this under a type 4 XA driver using the apache bench. This allow you to start a number of parallel threads executing the same application and by wrapping this in a script (Rexx or shell) you can also control how many loops you will conduct. We used this to verify the stability, the performance and the workload balancing in the product.

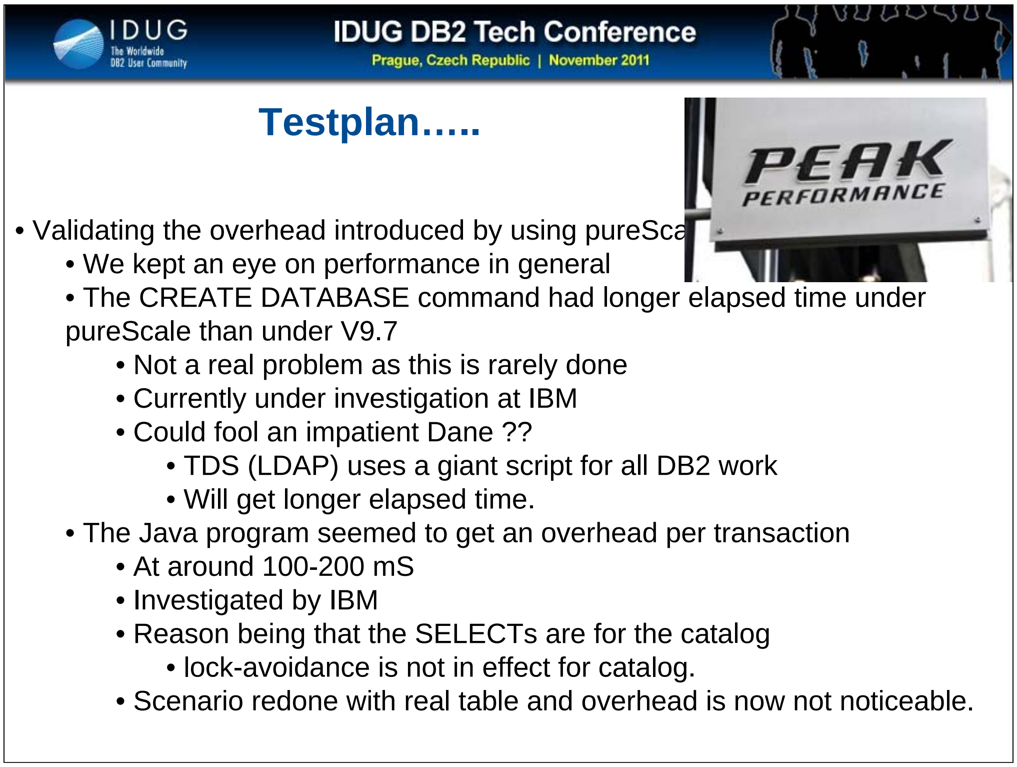
WAS data source WLB setting



□	currentFunctionPath		Specifies the value for the CURRENT FUNCTION PATH special register of the DB2 server.	false
□	cursorSensitivity		Specifies whether java.sql.ResultSet.TYPE_SCROLL_SENSITIVE maps to sensitive dynamic or sensitive static scroll. This property is ignored for insensitive scrollable cursors. The default is 0 (TYPE_SCROLL_SENSITIVE_STATIC). Alternatively, it can be set to 1 (TYPE_SCROLL_SENSITIVE_DYNAMIC) to have TYPE_SCROLL_SENSITIVE cursors map to sensitive dynamic cursors or set to 2 (TYPE_SCROLL_ASENSITIVE) to use SENSITIVE whenever possible, otherwise, fallback to INSENSITIVE when the request doesn't support SENSITIVE.	false
□	keepDynamic		Determines whether DB2 keeps dynamic SQL statements after commit points. This is applicable only when the target server is DB2 for z/OS. The property default depends on the server. Specifying 1 means YES, DB2 keeps dynamic SQL statements after commit points. Specifying 2 means NO, DB2 does not keep dynamic SQL statements after commit points. If this property is not set defaults are assumed. If the JDBC package set is custom bound without these defaults, this property must be set. The driver has no way to know how custom JDBC packages were bound under current DRDA architecture. This property is not a directive to server, but rather a driver directive.	false
□	currentLockTimeout		Directive to DB2 servers to either wait indefinitely for a lock or to wait for a certain amount of time for a lock when the lock cannot be obtained immediately. A value of zero means no wait. A value of -1 means to wait indefinitely. A positive integer indicates the number of seconds to wait for a lock.	false
□	useTransactionRedirect	false	Configure this property if your backend uses the Database Partitioning Feature (DPF) and your partitioning key remains constant throughout a transaction. This property is only valid when configuring the XA datasource with Type 4	false
□	clientRerouteAlternateServerName	aix189.bdunet.dk		false
■	clientRerouteAlternatePortNumber	50000		false
■	enableSysplexWLB	true		false

Page: 2 of 2 Total 29

To set the workload balancing at the transaction layer you need to set some properties in the WAS admin console under “custom properties”.



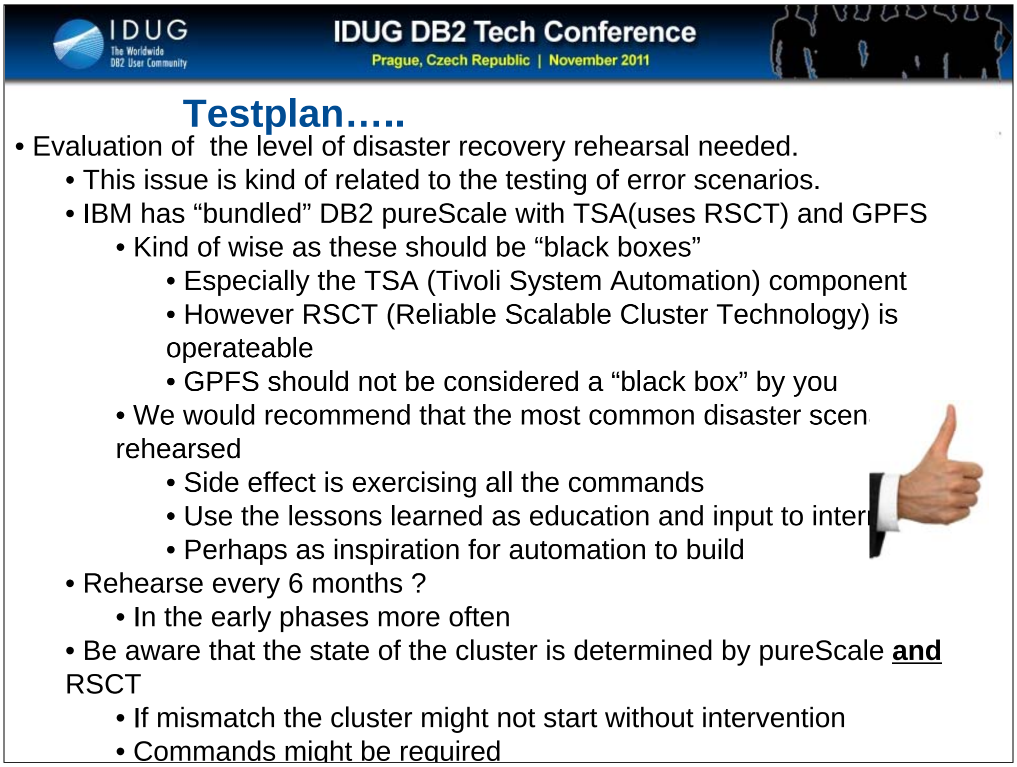
Testplan…..

PEAK
PERFORMANCE

- Validating the overhead introduced by using pureSca
  - We kept an eye on performance in general
  - The CREATE DATABASE command had longer elapsed time under pureScale than under V9.7
    - Not a real problem as this is rarely done
    - Currently under investigation at IBM
    - Could fool an impatient Dane ??
      - TDS (LDAP) uses a giant script for all DB2 work
      - Will get longer elapsed time.
  - The Java program seemed to get an overhead per transaction
    - At around 100-200 mS
    - Investigated by IBM
    - Reason being that the SELECTs are for the catalog
      - lock-avoidance is not in effect for catalog.
    - Scenario redone with real table and overhead is now not noticeable.

We were surprised that a common thing like a CREATE DATABASE took much longer time under pureScale than under DB2 UDB 9.7. As this is very rarely done the biggest danger in this is that users (installers) get impatient and cancel the process.

What worried us more was that we saw a big overhead on all transactions. IBM investigated and it proofed to be a consequence of the Java program selecting in the catalog where pureScale's lock-avoidance is not in effect. So properly/surely not a problem in a real world…..



IDUG DB2 Tech Conference

Prague, Czech Republic | November 2011

Testplan…..

- Evaluation of the level of disaster recovery rehearsal needed.
  - This issue is kind of related to the testing of error scenarios.
  - IBM has “bundled” DB2 pureScale with TSA(uses RSCT) and GPFS
    - Kind of wise as these should be “black boxes”
      - Especially the TSA (Tivoli System Automation) component
      - However RSCT (Reliable Scalable Cluster Technology) is operateable
      - GPFS should not be considered a “black box” by you
    - We would recommend that the most common disaster scen rehearsed
      - Side effect is exercising all the commands
      - Use the lessons learned as education and input to inter
      - Perhaps as inspiration for automation to build
  - Rehearse every 6 months ?
    - In the early phases more often
  - Be aware that the state of the cluster is determined by pureScale **and** RSCT
    - If mismatch the cluster might not start without intervention
    - Commands might be required

During the testing of all the possible and impossible error scenarios we saw that our knowledge of the underlying components GPFS, RSCT and TSA was not good enough as the health of the system involves decisions and actions in all these components. We advise new users of pureScale to invest time in getting familiar with these components and their commands. Understand that this is an investment because you might newer get in to a situation where this knowledge will be needed. But in a disaster knowledge can make the difference….

So please rehearse all the unthinkable scenarios in time of peace !!!

IDUG The Worldwide DB2 User Community

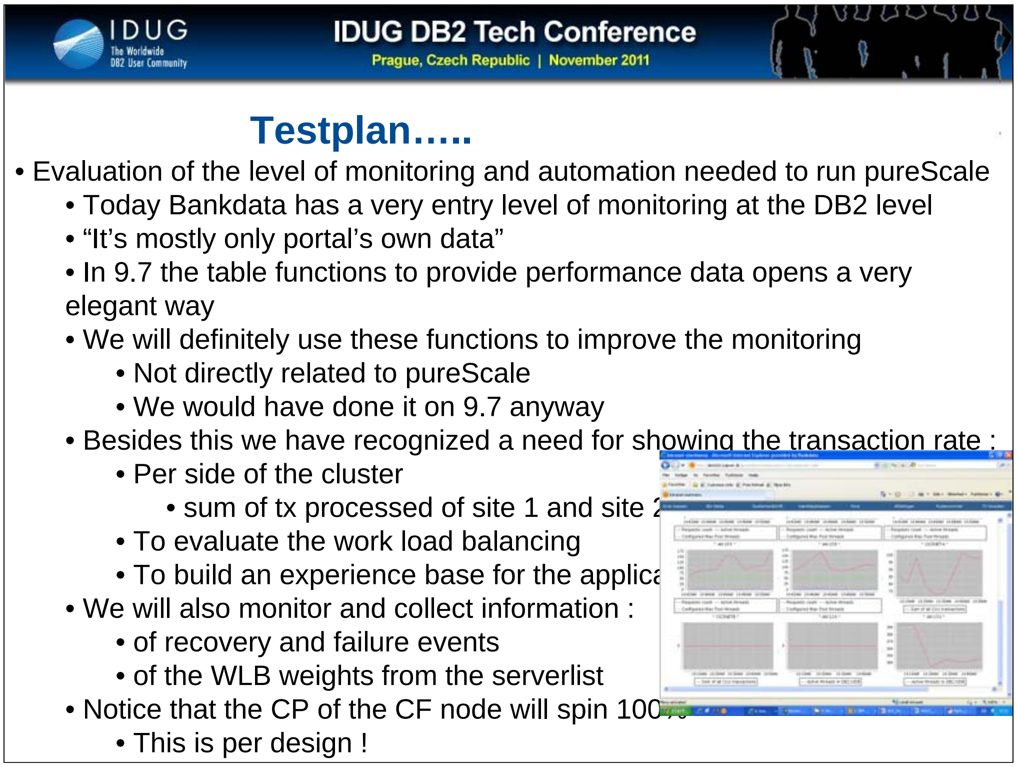
IDUG DB2 Tech Conference

Prague, Czech Republic | November 2011

Testplan…..

- Validation of the documentation level for the product
    - The general doc level of the product is considered to be at a high level
    - Some of the more technical documentation could be improved
        - For instance the relation/interaction of GPFS, TSA, RSCT and pureScale
        - Better walk through of commands to use in disaster situation
            - If the cluster will not start then….
    - Better description of how to move data into a pureScale cluster
        - Including a script to do the db2look and db2move
        - with the needed modifications
        - and that will find options that db2look does
    - It can be expected that this area will be improve

For a systems programmer this is the worst task of all : reading documentation.   We went through much of it and our feeling was that the base doc was well written and adequate but there was surely a need for more technical doc and white papers to help in understanding the internals in all the related components and how these interact with pureScale. This has been taken “ad notam” by IBM….

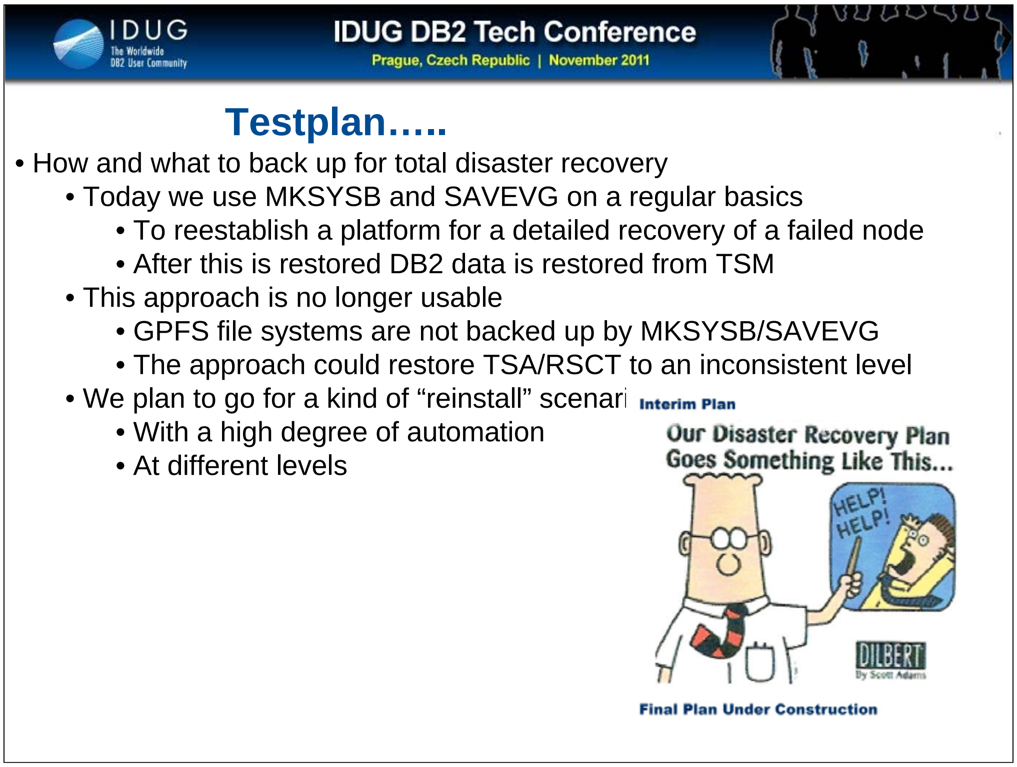


Testplan…..

- Evaluation of the level of monitoring and automation needed to run pureScale
  - Today Bankdata has a very entry level of monitoring at the DB2 level
  - “It’s mostly only portal’s own data”
  - In 9.7 the table functions to provide performance data opens a very elegant way
  - We will definitely use these functions to improve the monitoring
    - Not directly related to pureScale
    - We would have done it on 9.7 anyway
  - Besides this we have recognized a need for showing the transaction rate :
    - Per side of the cluster
      - sum of tx processed of site 1 and site 2
    - To evaluate the work load balancing
    - To build an experience base for the applica
  - We will also monitor and collect information :
    - of recovery and failure events
    - of the WLB weights from the serverlist
  - Notice that the CP of the CF node will spin 100%
    - This is per design !

When looking at our feelings concerning monitoring it is important to understand that we were running a DB2 UDB 9.1.x at that time. In 9.7 all these wonderful table functions arrived that will give you the power of seeing all the performance metrics in an easy way. Our monitoring today just graphs the transaction rates and some thread info with alerts case something goes wrong. If we decided NOT to go to pureScale we would go to DB2 UDB 9.7 and we would change to using all these metrics to modernize our monitoring of the DB2 UDB.

However if we look at our monitoring for our CICS on z/OS running SYSPLEX we simply show one plot with the sum of executed transactions on ALL the CICS regions running on each “side”. If a CICS region crashes it will not be reflected as long as the surviving members are capable of processing the incoming transactions and System Automation will simply restart the failing CICS. With pureScale we will do something similar : plot the database activity on a plot per “side” (read : computer center) and we will ensure to collect available information of events as member down and recovery events. We also expect to show information of the weights used by workload balancing.



Testplan…..

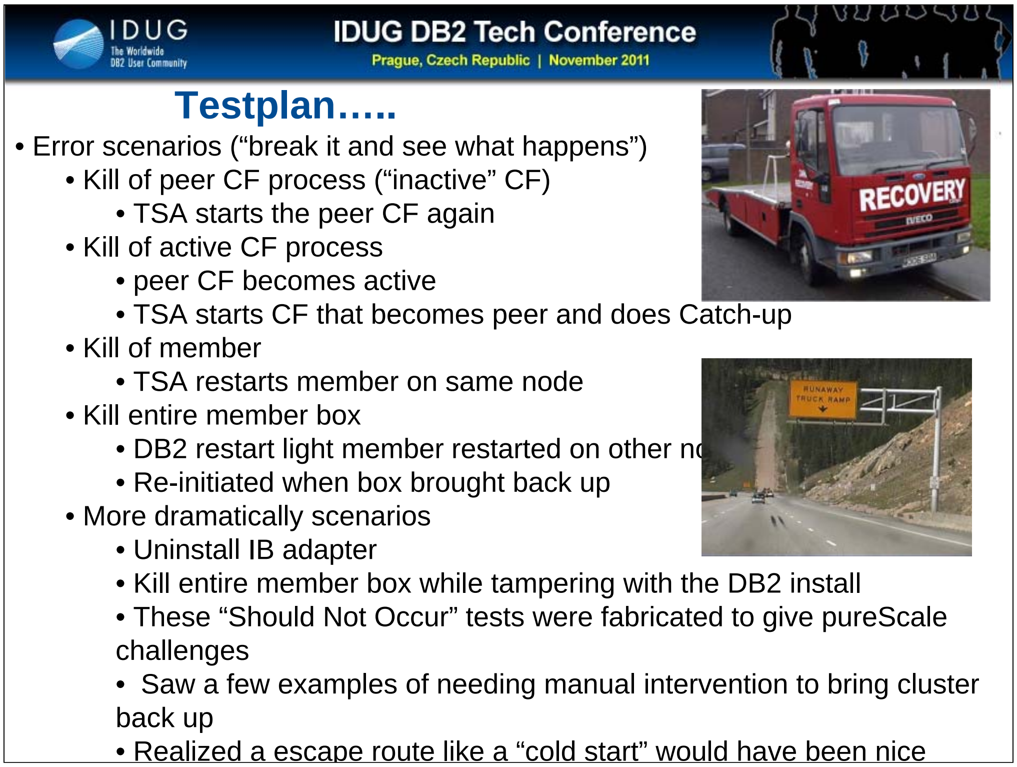
- How and what to back up for total disaster recovery
  - Today we use MKSYSB and SAVEVG on a regular basics
    - To reestablish a platform for a detailed recovery of a failed node
    - After this is restored DB2 data is restored from TSM
  - This approach is no longer usable
    - GPFS file systems are not backed up by MKSYSB/SAVEVG
    - The approach could restore TSA/RSCT to an inconsistent level
  - We plan to go for a kind of “reinstall” scenari
    - With a high degree of automation
    - At different levels

Interim Plan

Final Plan Under Construction

For the planning for the total disaster (loss of both centers) we today use a MKSYSB and SAVEVG to be able to rebuild the nodes in the applications.

However going to GPFS this is no longer a valid approach and the most likely strategy is to have some sort of reinstall plans and automation. This is one of the areas where a white paper will be very welcome….


RECOVERY
IVECO
RUNAWAY TRUCK RAMP

IDUG The Worldwide DB2 User Community

IDUG DB2 Tech Conference

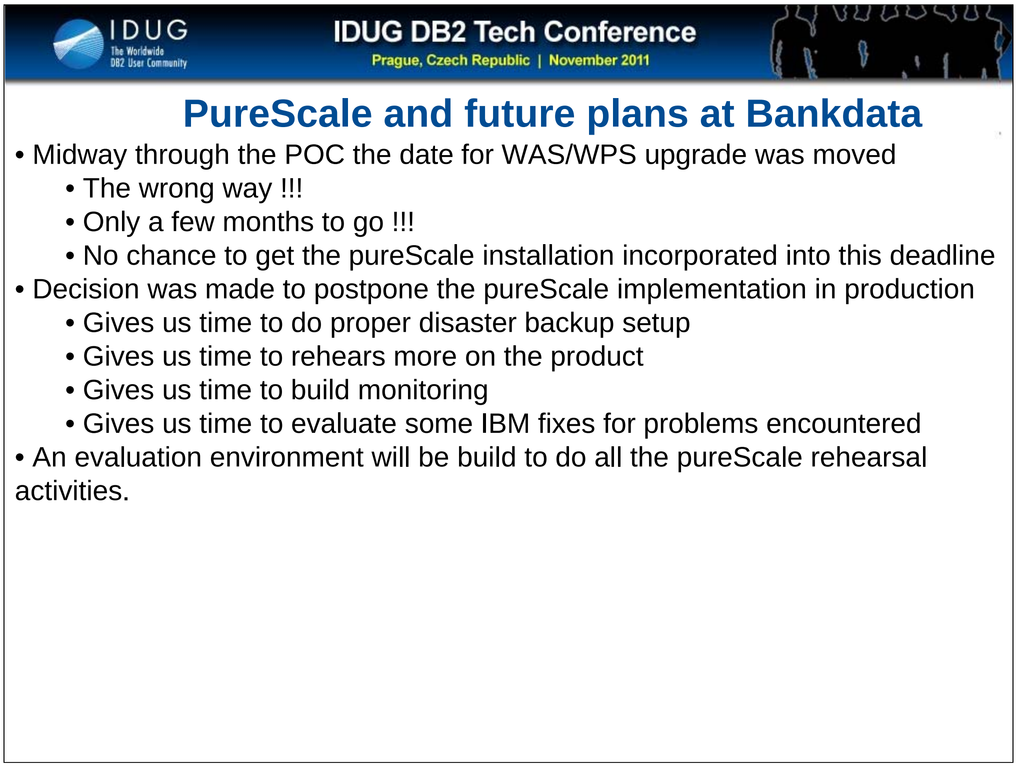
Prague, Czech Republic | November 2011

Testplan…..

- Error scenarios (“break it and see what happens”)
  - Kill of peer CF process (“inactive” CF)
    - TSA starts the peer CF again
  - Kill of active CF process
    - peer CF becomes active
    - TSA starts CF that becomes peer and does Catch-up
  - Kill of member
    - TSA restarts member on same node
  - Kill entire member box
    - DB2 restart light member restarted on other nc
    - Re-initiated when box brought back up
  - More dramatically scenarios
    - Uninstall IB adapter
    - Kill entire member box while tampering with the DB2 install
    - These “Should Not Occur” tests were fabricated to give pureScale challenges
    - Saw a few examples of needing manual intervention to bring cluster back up
    - Realized a escape route like a “cold start” would have been nice

The final heat : testing all the scenarios that is thinkable and unthinkable. We tested all the thinkable ones and these did exactly what they should with no or with smaller interruption in service.

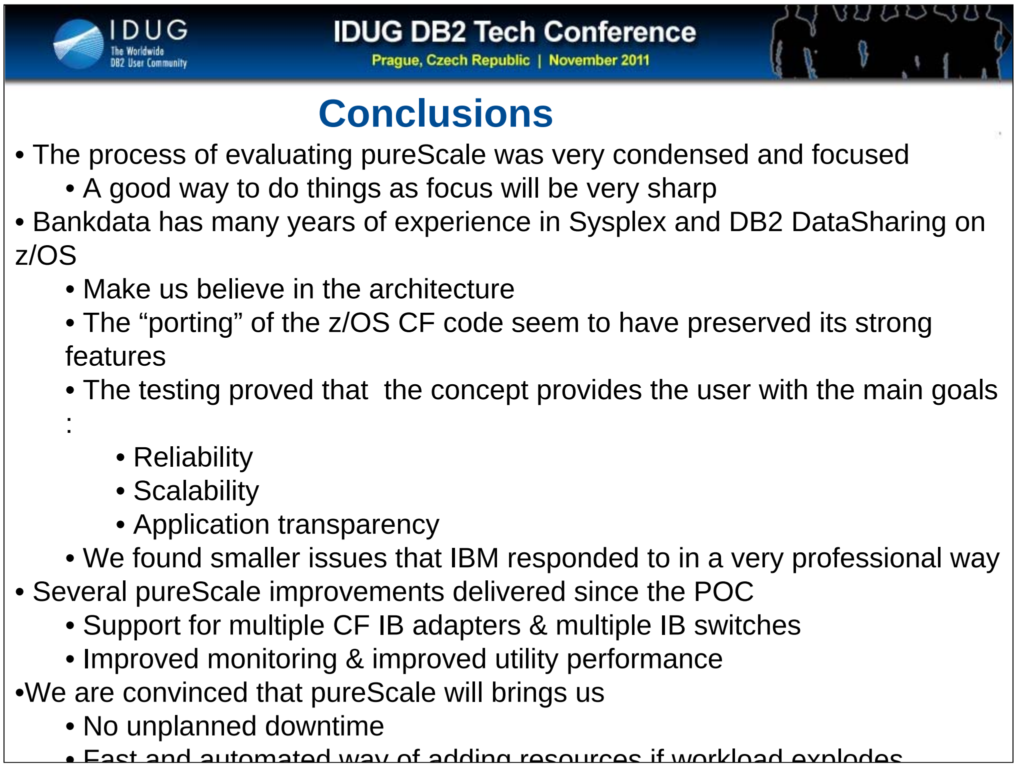
We also did things that will never happen and some of these indeed could bring the system in a state where it could not start automatically. IBM has been very aggressive in finding ways to handle these situations as well inside the code even though is was “ShouldNotOccur” situations.



PureScale and future plans at Bankdata

- Midway through the POC the date for WAS/WPS upgrade was moved
  - The wrong way !!!
  - Only a few months to go !!!
  - No chance to get the pureScale installation incorporated into this deadline
- Decision was made to postpone the pureScale implementation in production
  - Gives us time to do proper disaster backup setup
  - Gives us time to rehears more on the product
  - Gives us time to build monitoring
  - Gives us time to evaluate some IBM fixes for problems encountered
- An evaluation environment will be build to do all the pureScale rehearsal activities.

Where are we today ? In the Bankdata plans for upgrading these 2 vital application there were included enough slack to do a migration to pureScale. However these plans were changed so the applications needed to be upgraded earlier that expected. We therefore had to postpone the migration to pureScale and are at the moment using this extra time to get even more familiar with the product and verify some of the improvements to come.



IDUG The Worldwide DB2 User Community

IDUG DB2 Tech Conference

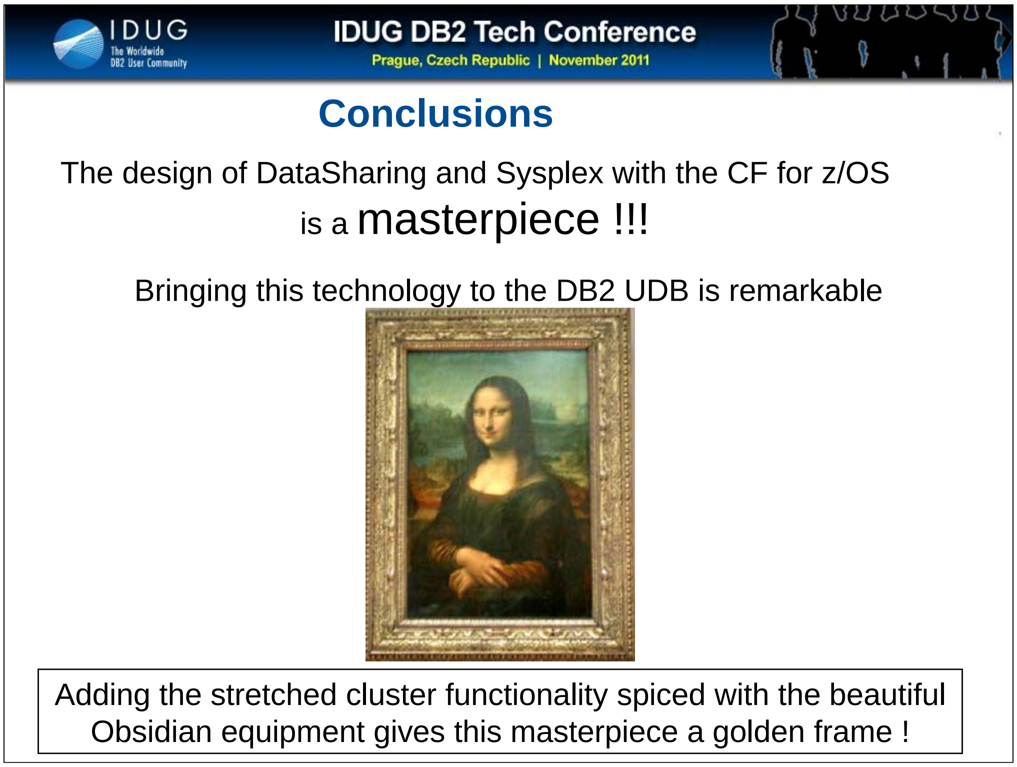
Prague, Czech Republic | November 2011

Conclusions

- The process of evaluating pureScale was very condensed and focused
  - A good way to do things as focus will be very sharp
- Bankdata has many years of experience in Sysplex and DB2 DataSharing on z/OS
  - Make us believe in the architecture
  - The “porting” of the z/OS CF code seem to have preserved its strong features
  - The testing proved that the concept provides the user with the main goals :
    - Reliability
    - Scalability
    - Application transparency
  - We found smaller issues that IBM responded to in a very professional way
- Several pureScale improvements delivered since the POC
  - Support for multiple CF IB adapters & multiple IB switches
  - Improved monitoring & improved utility performance
- We are convinced that pureScale will brings us
  - No unplanned downtime
  - Fast and automated way of adding resources if workload explodes

The conclusion is that the pureScale product is working excellent and that it is working very similar to DataSharing on z/OS as it is to some degree a porting of the code and ideas from this to the DB2 UDB.

We met a very dedicated and skilled staff at the IBM Toronto Lab that responded quickly and competent on our problems and silly questions.



Conclusions

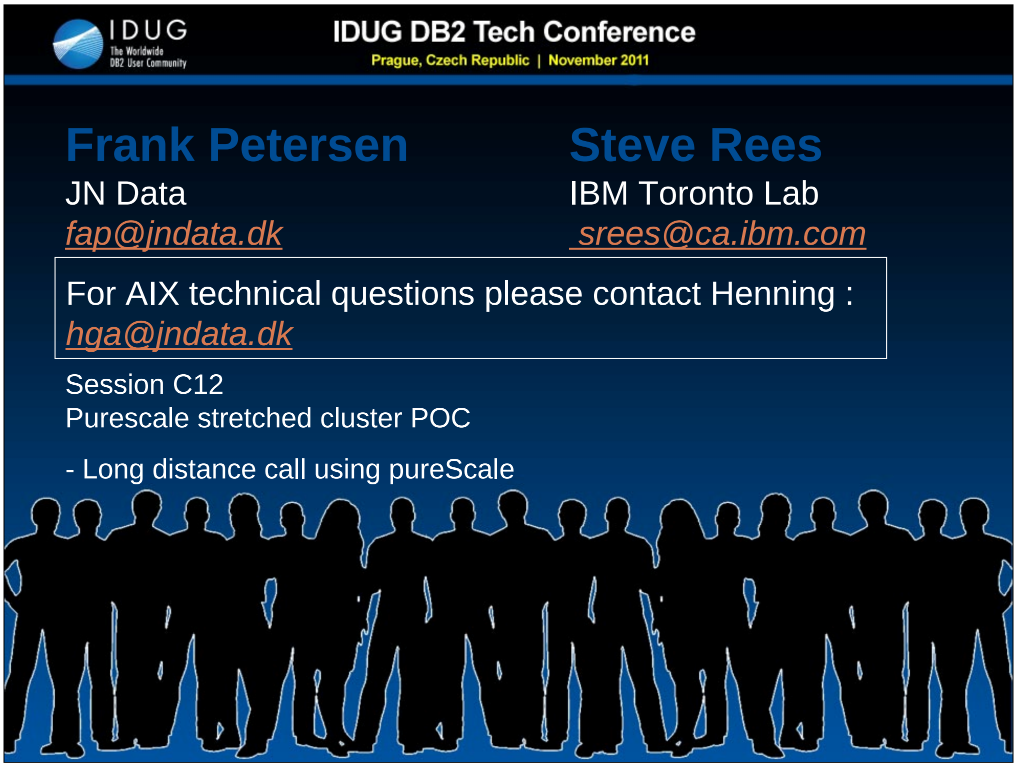
The design of DataSharing and Sysplex with the CF for z/OS
is a masterpiece !!!

Bringing this technology to the DB2 UDB is remarkable

Adding the stretched cluster functionality spiced with the beautiful Obsidian equipment gives this masterpiece a golden frame !

The conclusion is that the pureScale product is working excellent and that it is working very similar to DataSharing on z/OS as it is to some degree a porting of the code and ideas from this to the DB2 UDB.

We met a very dedicated and skilled staff at the IBM Toronto Lab that responded quickly and competent on our problems and silly questions.



IDUG DB2 Tech Conference

Prague, Czech Republic | November 2011

Frank Petersen
JN Data
fap@jndata.dk

Steve Rees
IBM Toronto Lab
srees@ca.ibm.com

For AIX technical questions please contact Henning : *hga@jndata.dk*

Session C12
Purescale stretched cluster POC

- Long distance call using pureScale