Db2 pureScale enhancements in V11.5

Dr Toby Haynes PhD

IBM Canada Ltd

IDUG

Leading the Db2 User Community since 1988

Please note :

- IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM's sole discretion.
- Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.
- The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.
- The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
- Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Notices and disclaimers

- © 2019 International Business Machines Corporation. No part of this document may be reproduced or transmitted in any form without written permission from IBM.
- **U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.**
- Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.
- IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."
- **Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.**
- Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.
- References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.
- Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.
- It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

New developments in Db2 pureScale for v11.5 and beyond

- Highlight the many areas of development for Db2 pureScale.
- Discuss how Db2 pureScale is being shaped by customer needs and requests.

Overview

- For Db2 pureScale v11.5 GA we've targeted improvements in
  - Security
  - Consistency
  - Robustness
  - Performance
- The majority of work done is in response to **user feedback**

Security (1|2)

- Security is needed everywhere
  - Application security
  - Network security
  - Authorization
  - Encryption
  - … and more

Security (2|2)

- Db2 has a hardened code base (including AppScan static code analysis)
- Also offers options for
  - Encrypted connections through SSH
  - Encrypted database
- … and now Db2 v11.5 supports host-based firewalls

Consistency

- Things that “just work” across the cluster
  - Less manual setup
  - More consistent behaviour
    - Address features that behave subtly differently between standalone, HADR and pureScale configurations

Robustness

- More “awareness” built into the cluster
  - Automatic configuration of public ethernet devices
    - Enabled for GDPC too!
  - Automatic kernel module compilation for IBM Spectrum Scale
- Updates to the software stack
  - IBM Spectrum Scale
  - Tivoli System Automation Multi-Platform (TSA)
  - IBM Reliable Scalable Cluster Technologies (RSCT)
- Even more testing before release

Performance

- Almost 10 years since Db2 pureScale was announced in October 2009
  - POWER6 maxed out (P595) at 32 sockets, 2 cores per socket, 2 hardware threads per core
    - 128 logical CPUs
  - Today's POWER9 largest configuration (E980) has 16 sockets, 12 cores per socket, 8 hardware threads per core
    - 1536 logical CPUs
- Db2 pureScale v11.5 can make use of more processors and new hardware

Better Security

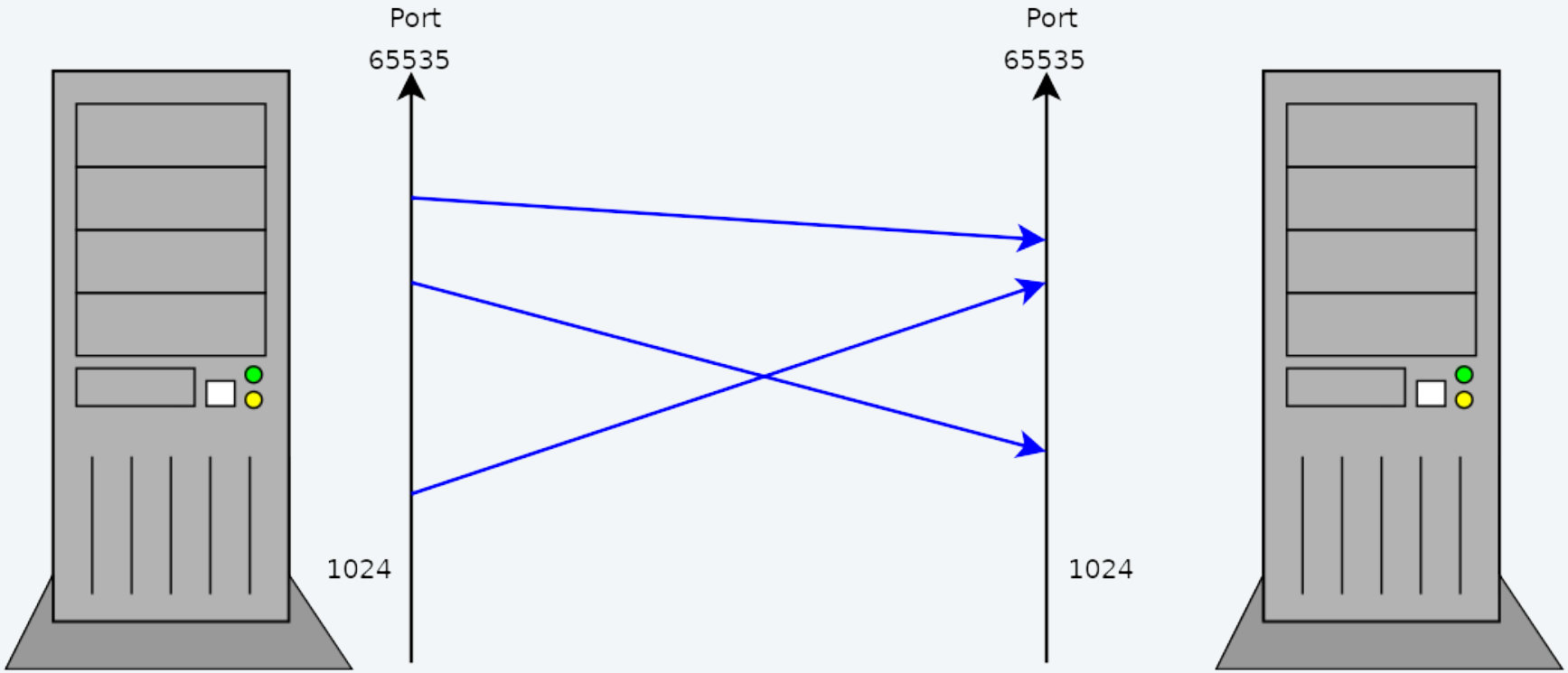
Support for host-based firewalls (1|8)

- Prior to v11.5, enabling a firewall on any host within the cluster would stop normal operations
- Host-based firewalls attempt to limit access, e.g. to a whitelist of
  - ports
  - hosts
  - related connections

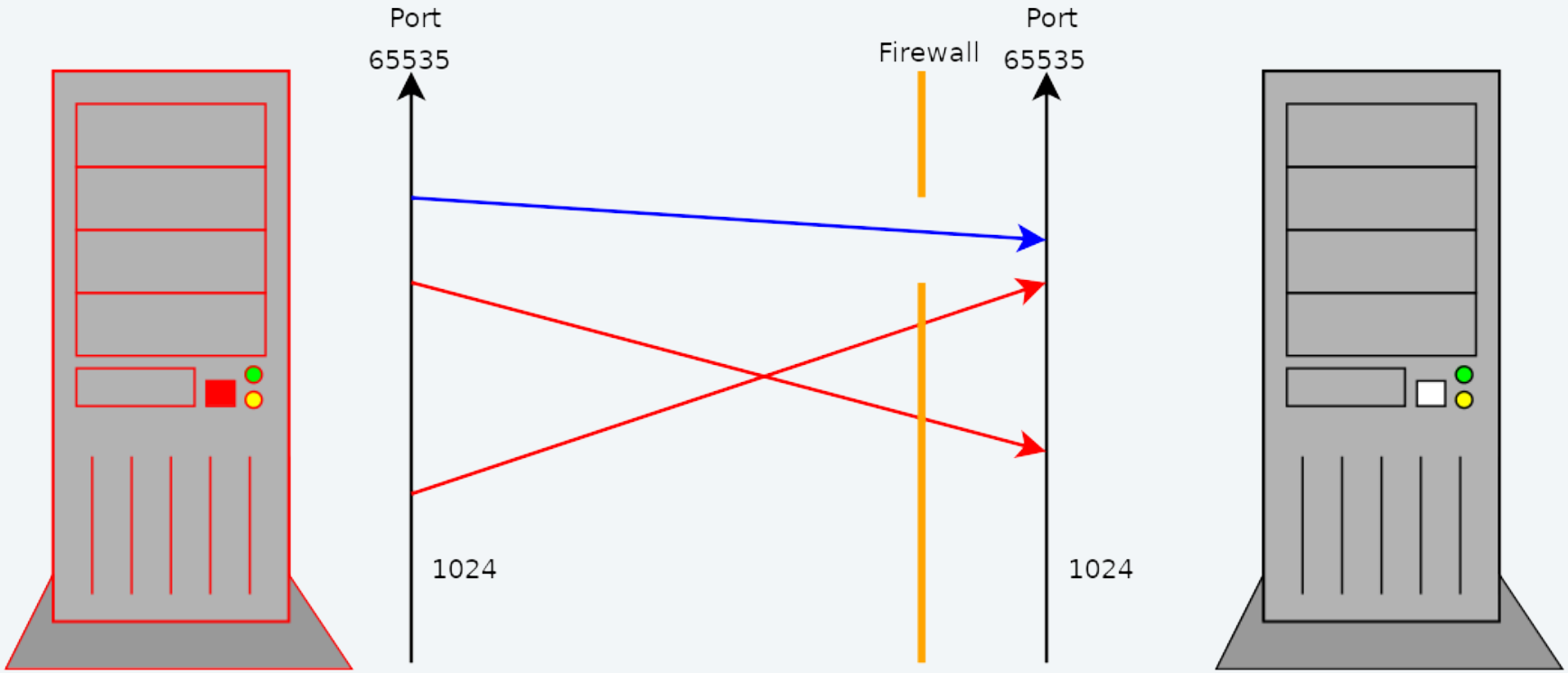
Support for host-based firewalls (2|8)

- **The Problem**
  - Many Db2 operations, across install, upgrade and regular operation, make connections between hosts.
  - These connections use ports across much of the 1024 – 65535 port range
  - Can’t predict the usage patterns of a running server

Support for host-based firewalls (3|8)



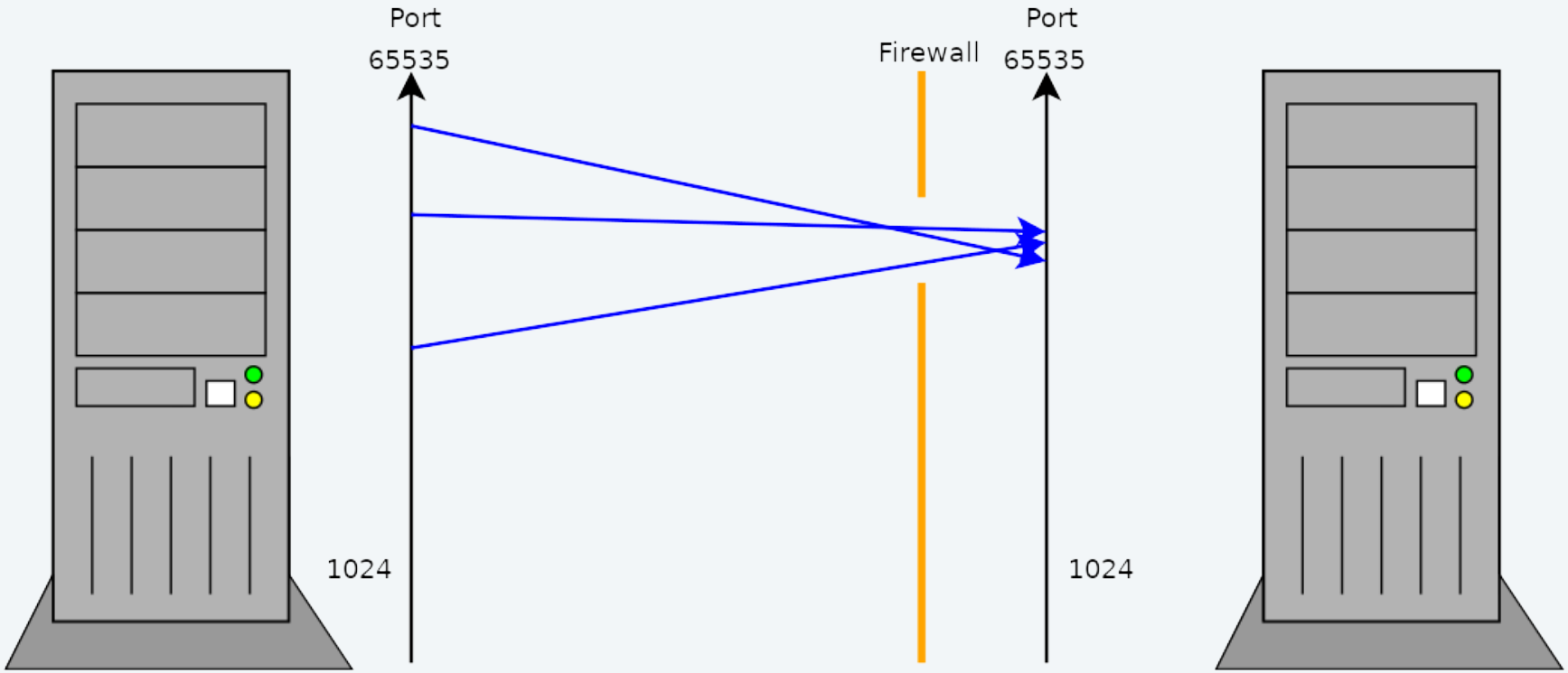
Support for host-based firewalls (4|8)



Support for host-based firewalls (5|8)

- One setting and DONE
  - In Db2 v11.5, users can specify a port range for which will be used for all Db2 activities, from before install throughout deployment and during production.
- Configurable options
  - `db2set DB2_FIREWALL_PORT_RANGE=<StartPort>-<EndPort>`

Support for host-based firewalls (6|8)



Support for host-based firewalls (7|8)

- IBM Spectrum Scale integrated
  - The Db2 core engine is not the only part of Db2 pureScale
  - IBM Spectrum Scale also makes connections between hosts.
- Part of the port range specified for DB2_FIREWALL_PORT_RANGE is automatically configured for Spectrum Scale.

Support for host-based firewalls (8|8)

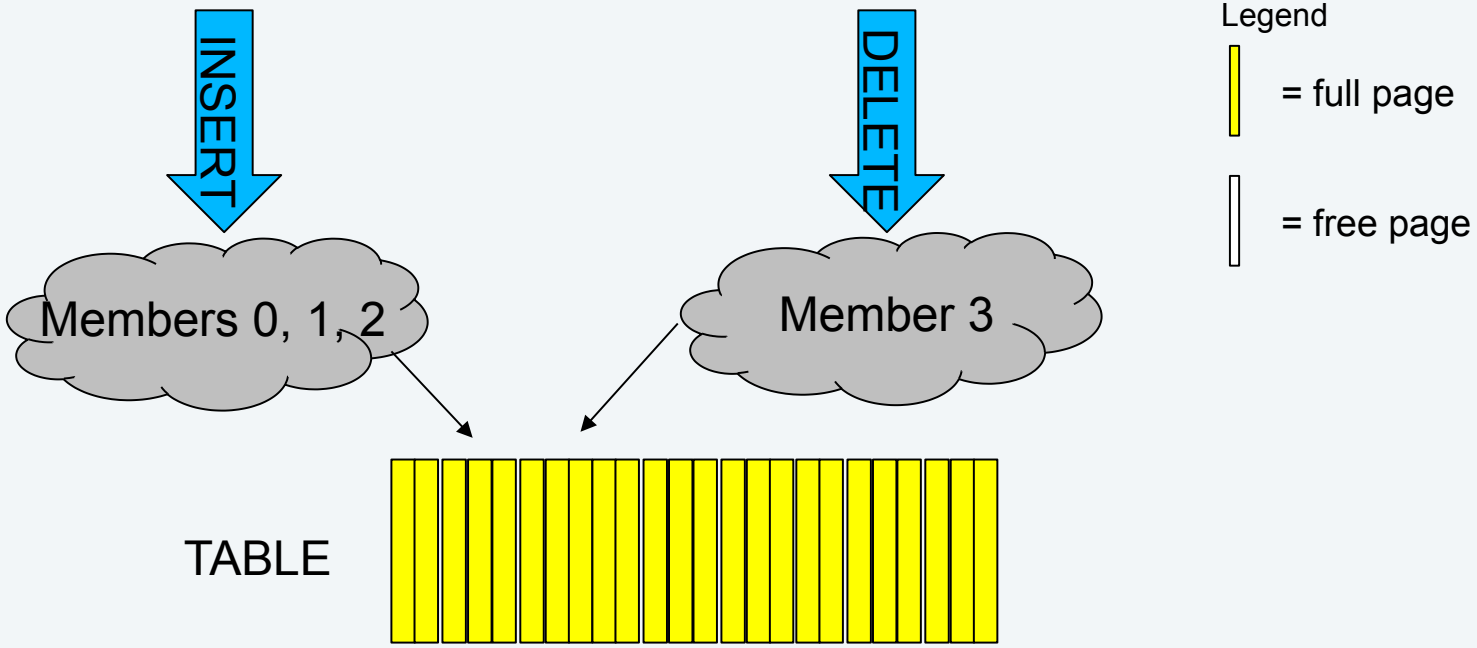
- User-managed Spectrum Scale
  - For customers who have existing IBM Spectrum Scale installations prior to Db2 pureScale, or who have already configured a port range, no change is made.

Consistency

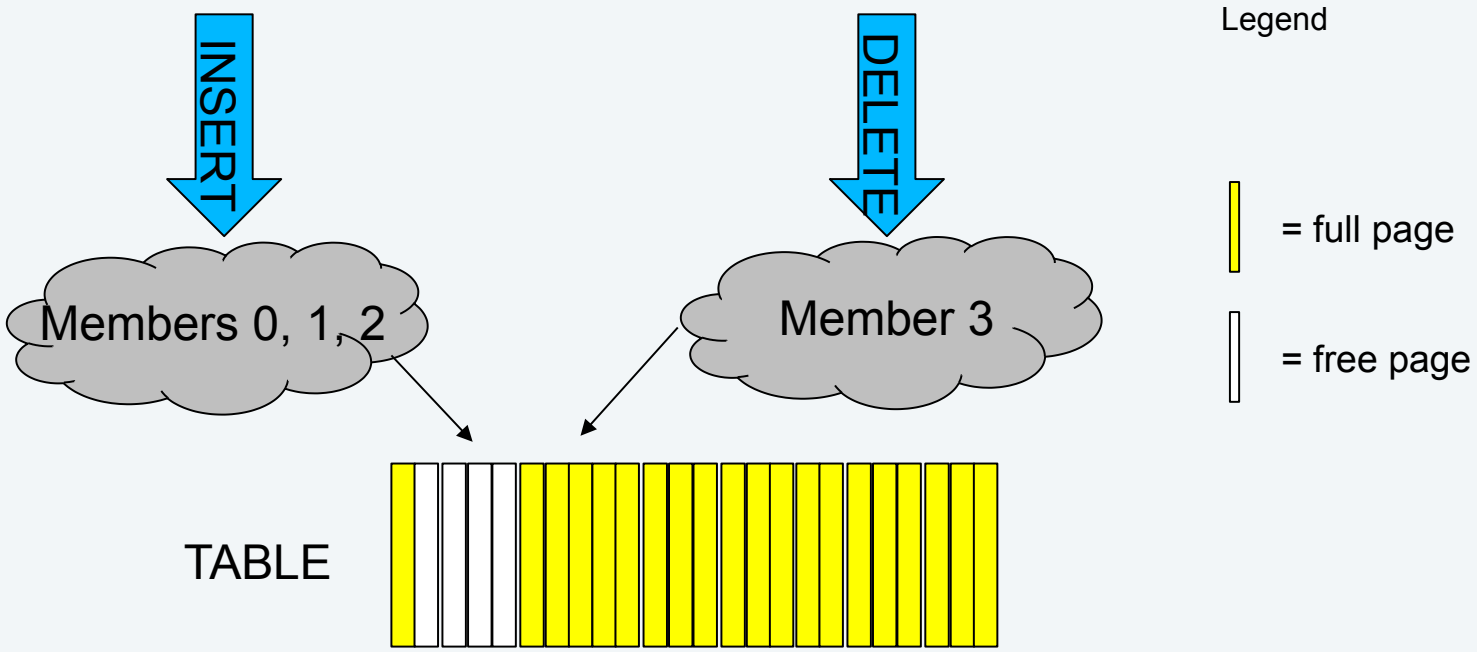
Improved cluster wide free space management (1|9)

- **Problem to solve**:
  - Excessive free pages in table when
    - One workload which exclusively consists of inserts to a table against a ***subset*** of members in the cluster
    - Another workload which consists of selects/updates/deletes to the same against **another subset** of members in the cluster,
- **Observed**:
  - Free pages in the table are not re-used as expected
  - Table scans on these tables take longer than needed
- Requested by multiple customers:
  - RFE 33189

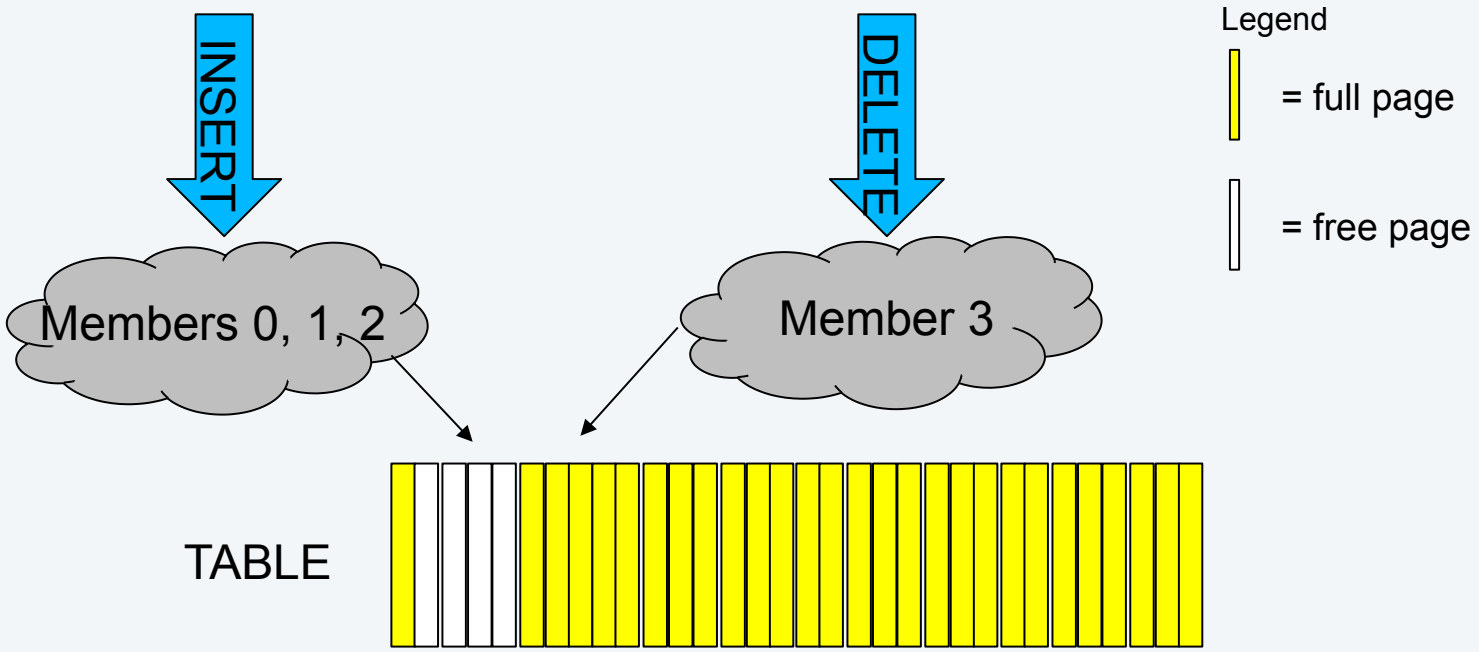
Improved cluster wide free space management (2|9)



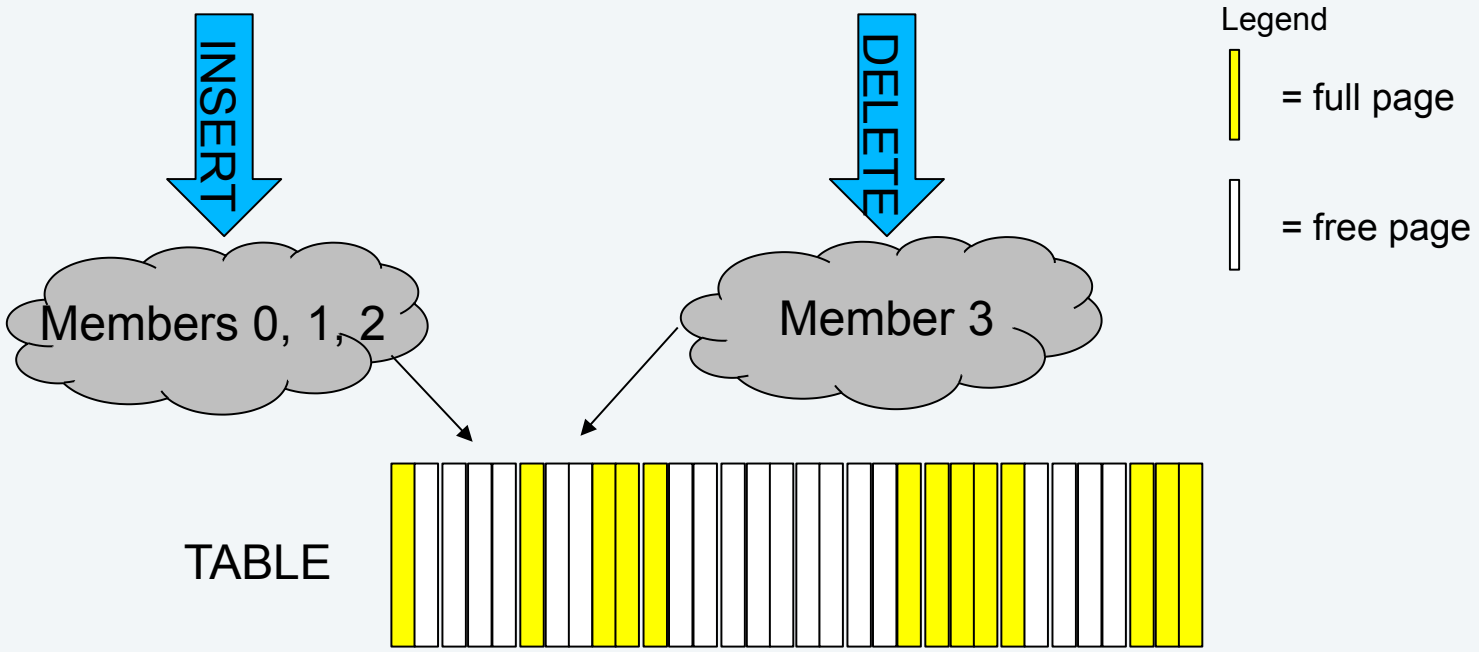
Improved cluster wide free space management (3|9)



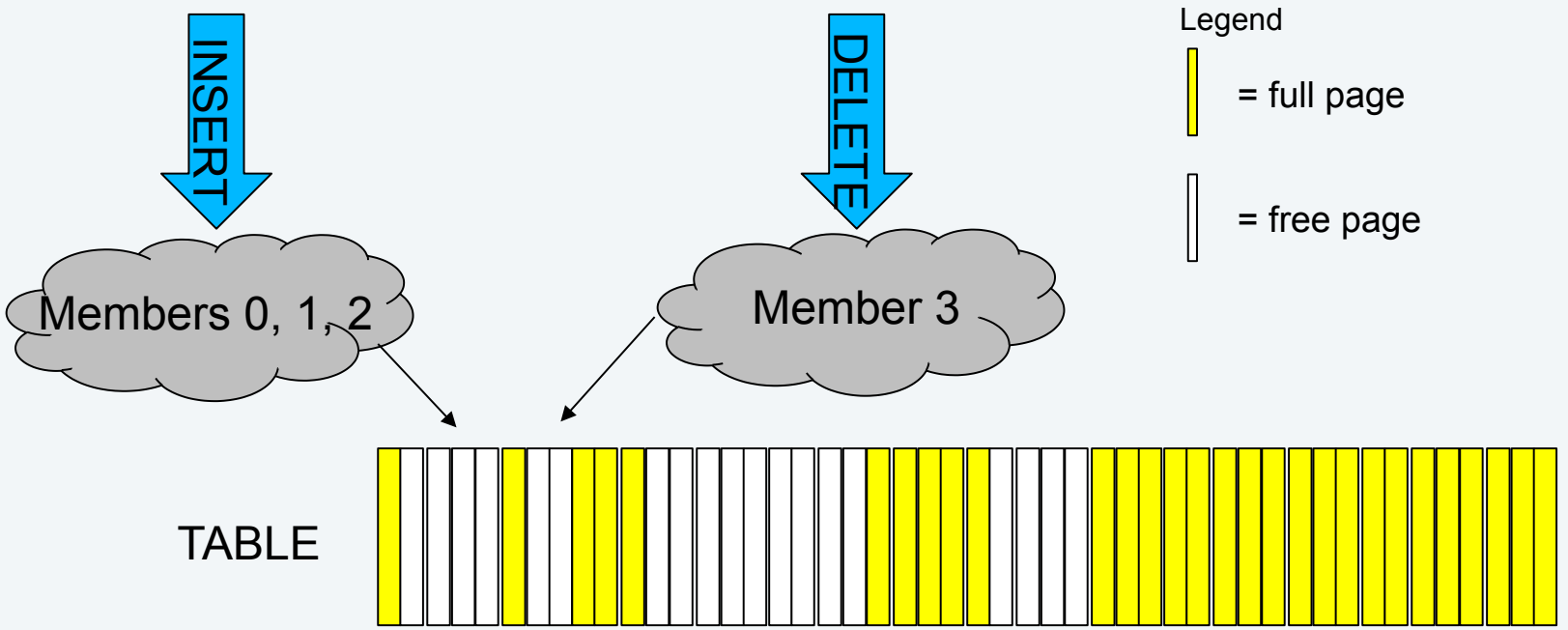
Improved cluster wide free space management (4|9)



Improved cluster wide free space management (5|9)



Improved cluster wide free space management (6|9)



Improved cluster wide free space management (7|9)

- Previous design
  - Each member tracks its own free space
  - When work load balancing is distributing the work around all the members
    - sooner or later a delete or update will indicate some new free space for this member

Improved cluster wide free space management (8|9)

- Member subsets debuted in v10.5
  - Separate applications into a subset of members – e.g.
    - Batch on two members
    - OLTP on two members
    - Analytics and clean up on two members
- Batch work may not drive any deletes and may not know about free space changes occurring during inserts.

Improved cluster wide free space management (9|9)

- Solution
  - The free space indicator has been moved to the CF
    - All the members can check for new free space efficiently and reliably
  - This is the **new default behavior** starting in Db2 v11.5 GA.
- The following registry variable can be set to revert back to old behavior of free space searching:

```
db2set DB2_SD_DISABLE_GLOBAL_SPACE_SEARCH=TRUE
```

Currently Committed

- CS isolation can cause lock timeouts and deadlocks with row-level locking
- *Currently Committed* semantics allow only committed data to be returned
  - Readers do NOT wait for writers to release row locks
  - Readers get data based on the data prior to the start of the write operation
- Debuted in v9.7

Currently Committed in pureScale v11.1 and earlier

- In v11.1, *currently committed* semantics only apply to lock conflicts between applications running on the same member.
  - i.e. lock conflicts with an application on a remote member have to wait for the lock to be released before processing the row.
- Only useful for applications with a member affinity or single member subset

Currently Committed in pureScale v11.5

- Now the last committed version of the row is retrieved from the remote member
  - Data is retrieved from the log buffer – no extra disk I/O is required
- Reduced lock wait performance!

Robustness

Automatic configuration of public network monitoring

- Goals:
  - Simplify configuration of GDPC
  - Provide consistent CF resources setup on all pureScale configurations
  - Automatically create public ethernet condition response pair for GDPC (during setup, repair and enter/exit maintenance)
- Result:
  - **8% faster recovery for public ethernet failure tests**

Automatic configuration of public network monitoring

Manual commands for GDPC	V11.1	V11.5
After typical install	33	0
After every maintenance operation	6	0

IBM Spectrum Scale (1|2)

- New major release – IBM Spectrum Scale version 5.0.2
- Usability:
  - Automatic Linux kernel module compilation enabled by default
  - Changes to existing Linux kernel levels will no longer require a manual compilation step
- Performance:
  - File system re-balancing is faster
  - concurrent updates to small shared directories are faster
  - small sub-block sizes improves I/O performance.

IBM Spectrum Scale (2|2)

- New configuration defaults
  - cipherList AUTHONLY
    - TLS handshaking is used to authenticate the hosts – communication between hosts is not encrypted or validated
  - autoBuildGPL yes
    - automatic linux kernel module compilation, for new upgraded kernel levels or upgraded Spectrum Scale
  - WorkerThreads 512
    - Replaces the deprecated worker1Threads/worker3Threads with single parameter
  - MaxFilesToCache 15000
    - Bump up under guidance from the Spectrum Scale team for better caching

Performance

Faster LOAD for range-partitioned tables (1|2)

- Multiple RFEs for faster LOAD performance
  - Especially for range partitioned tables
- The Problem:
  - LOAD requires the bufferpool to be flushed/purged of the target table data before it starts
  - In standalone installations (ESE), all these bufferpool operations are grouped together
  - In pureScale v11.1 and earlier, the flush/purge was done individually for each partition, each index and LOB columns
    - Required RPC to each member

Faster LOAD for range-partitioned tables (2|2)

- Improvements:
  - Grouped flush/purge for partitions, indexes and LOB columns
  - Result: 2x faster LOAD
- Further work to be done!

Multiple XI connections for RDMA (1|2)

- Cross Invalidation (XI) is the process that indicates that a change has been made to a page in a bufferpool
  - Many XI requests can be serialized on a single connection
  - More Inserts/Updates/Deletes drive more XI requests
- Multiple XI connections for TCP/IP sockets was delivered in Db2 v11.1 modpack 3 fixpack 3
  - Performance improvements for write-heavy workloads

Multiple XI connections for RDMA (2|2)

- Db2 pureScale v11.5 delivers support for multiple XI connections for RDMA
  - RoCE
  - Infiniband
- Some performance benefits even on “local” clusters
  - 1-2% improvement in transaction rates at 2 XI connections per member
  - More important to GDPC clusters using RoCE technologies

Support for 96 CF worker threads

- Db2 pureScale v11.1 supports up to 31 CF worker threads
  - Rule of thumb: Total number of member logical CPUs = number of worker threads * 12
    - e.g. 4 members with 96 logical CPUs would saturate 31 CF worker threads
  - CF_NUM_WORKERS
- Db2 pureScale v11.5 supports up to 96 CF worker threads
  - RDMA-based configurations must have at least (CF_NUM_WORKERS + 1) logical CPUs
  - TCP/IP interconnects are still recommended to use 8*(logical CPUs)

Better performing page cleaning for busy clusters

- The Problem:
  - Db2 v11.1 and earlier used a single Castout Class used in the Shared Communication Area
  - Busy page cleaning could bottleneck on the Castout Class latch
- The Solution:
  - 1024 Castout Classes now used by default
- Performance improvements for intensive OLTP workloads

What's next?

- Db2 pureScale in the Cloud,
  - Docker/ Containers/ Kubernetes
- Better, simpler verification of GDPC clusters
  - Base infrastructure landed in v11.5 – more work to do
- More validation of system configuration
  - /etc/hosts, /etc/rdma/dat.conf and more
- More performance work
  - LOAD with Range Partitioned Tables
- Hang detection for stuck member

What do YOU want to see?

- IBM Data and AI publishes a Roadmap on Aha!
  - https://ibm-analytics-roadmaps.mybluemix.net/
- Ideas can be proposed at
  - https://ibm-data-and-ai.ideas.aha.io/

Dr. Toby Haynes PhD
IBM Canada Ltd
thaynes@ca.ibm.com

IDUG
Leading the Db2 User Community since 1988