OPENFABRICS ALLIANCE

DB2 *pureScale*

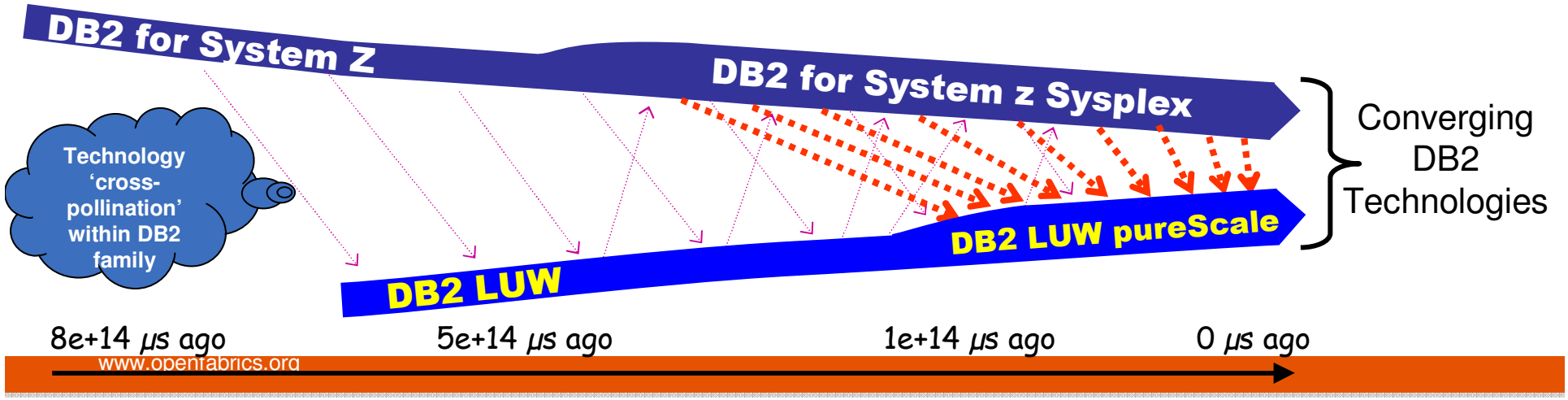
Steve Rees – IBM Canada Ltd.
March 15, 2010

Agenda

- Introduction to pureScale
  - 10,000 ft view
  - Goals & value for users
  - Technology overview
- Key Concepts & Internals
  - Major components & moving parts
  - Efficient scaling over low-latency interconnect
- Interconnect issues
  - Requirements, futures, etc.

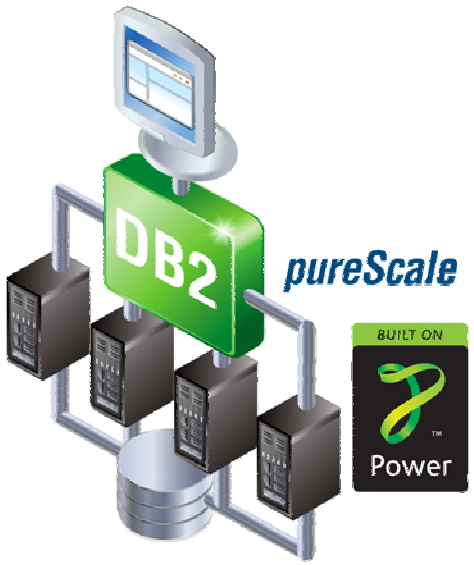
DB2 pureScale evolution from ~~10,000 ft~~ 3×10^{12} nm

- Before *pureScale*: DB2 for Linux/Unix/Windows (LUW) available in
  - SMP configurations
  - Shared-nothing clusters
- DB2 on the mainframe
  - SMP configurations
  - Shared-data clusters with System z Sysplex
  - Very well-known for continuous availability and OLTP scalability
- DB2 *pureScale* merges interconnect and database technologies to create scale-out, highly available database clusters



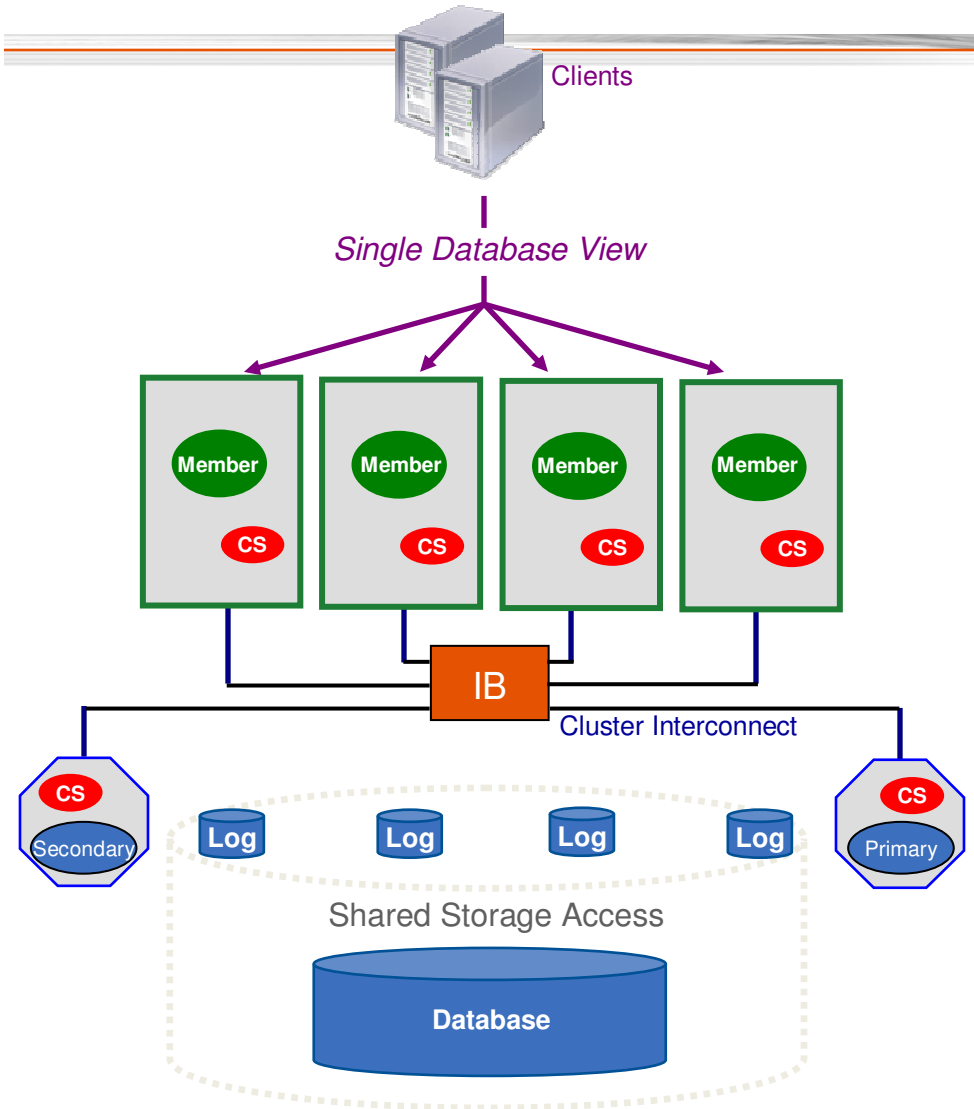
DB2 pureScale : Goals

- Unlimited Capacity
  - Any transaction processing or ERP workload
  - Start small
  - Grow easily
- Application Transparency
  - Avoids the risk and cost of tuning applications to the database topology
- Continuous Availability
  - Maintain service across planned and unplanned events



DB2 pureScale : Technology Overview

Leverage System z Sysplex Experience and Know-How



Clients connect anywhere and see a 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 machines

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

Low latency, high speed interconnect

- Special optimizations provide significant advantages on RDMA-capable interconnects (eg. Infiniband)

PowerHA pureScale technology

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

Data sharing architecture

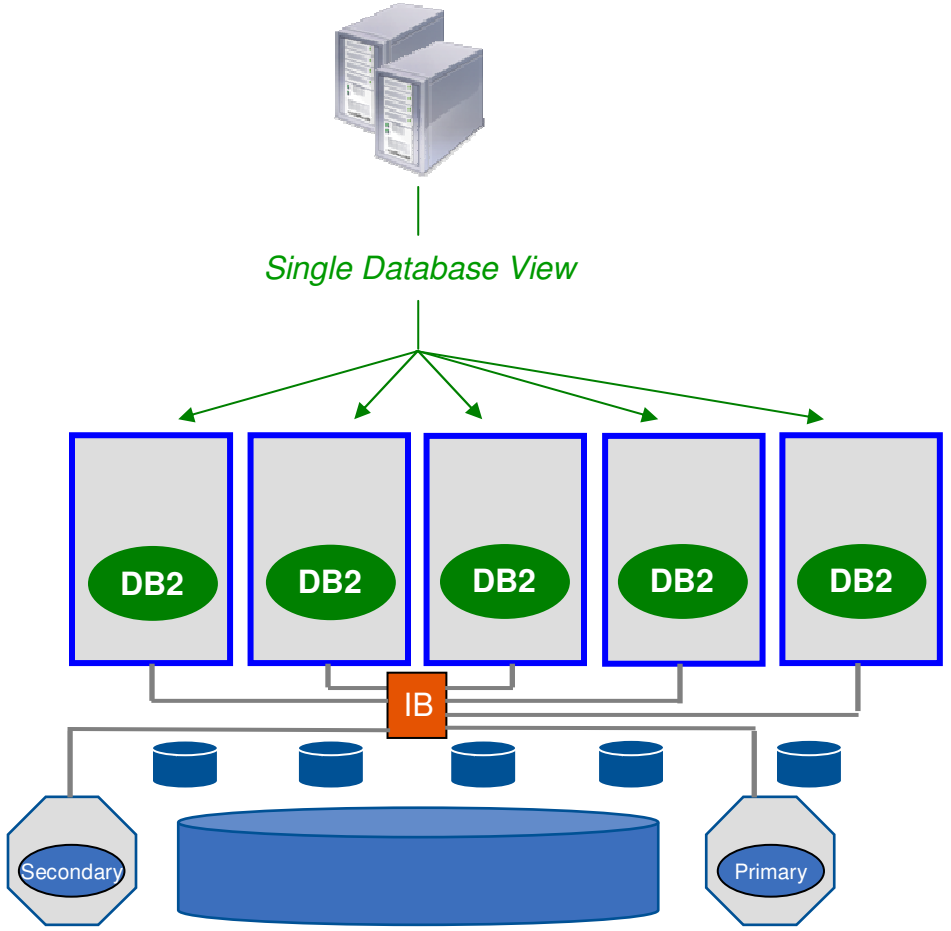
- Shared access to database
- Members write to their own logs
- Logs accessible from another host (used during recovery)

Integrated cluster services

- Failure detection, recovery automation (TSA / RSCT)
- Cluster file system (GPFS)

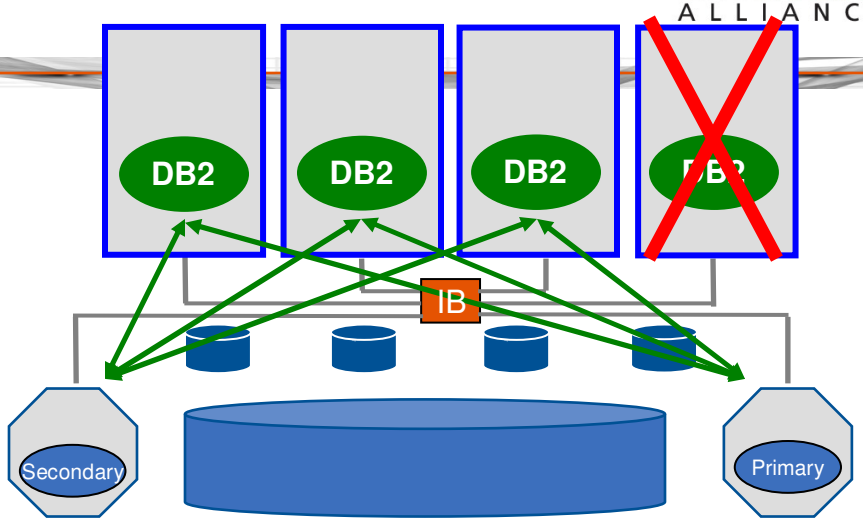
Easy scale-out

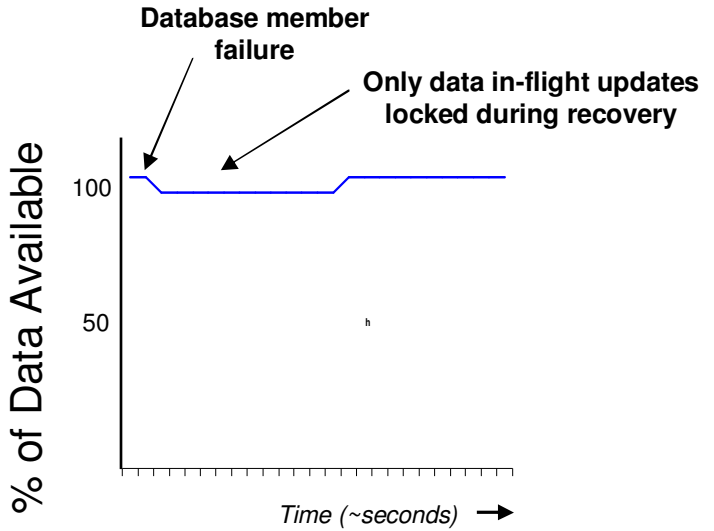
- To 128 members in initial release
  - High IOPS & BW of interconnect make this possible
- Efficient coherency protocols exploit low-latency interconnect to scale without application change
- Applications' workload automatically and transparently balanced across members
- No data redistribution required



Online Recovery

- A key pureScale design point is to maximize availability **during** failure recovery processing
- When a database member fails, only data *in-flight* on the failed member remains locked during the automated recovery
- High-speed interconnect means duration of even partial inaccessibility is limited





Agenda

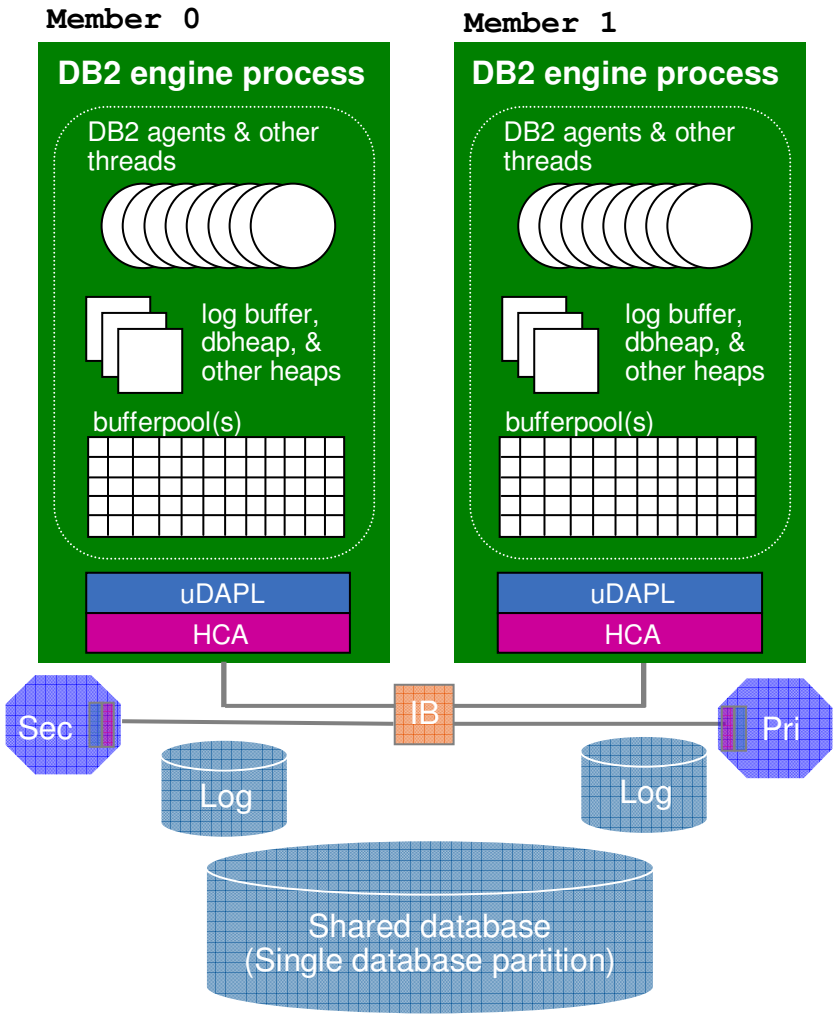
- Introduction to pureScale
  - 10,000 ft view
  - Goals & value for users
  - Technology overview

→ Key Concepts & Internals
  - Major components & moving parts
  - Efficient scaling over low-latency interconnect

- Interconnect issues
  - Requirements, futures, etc.

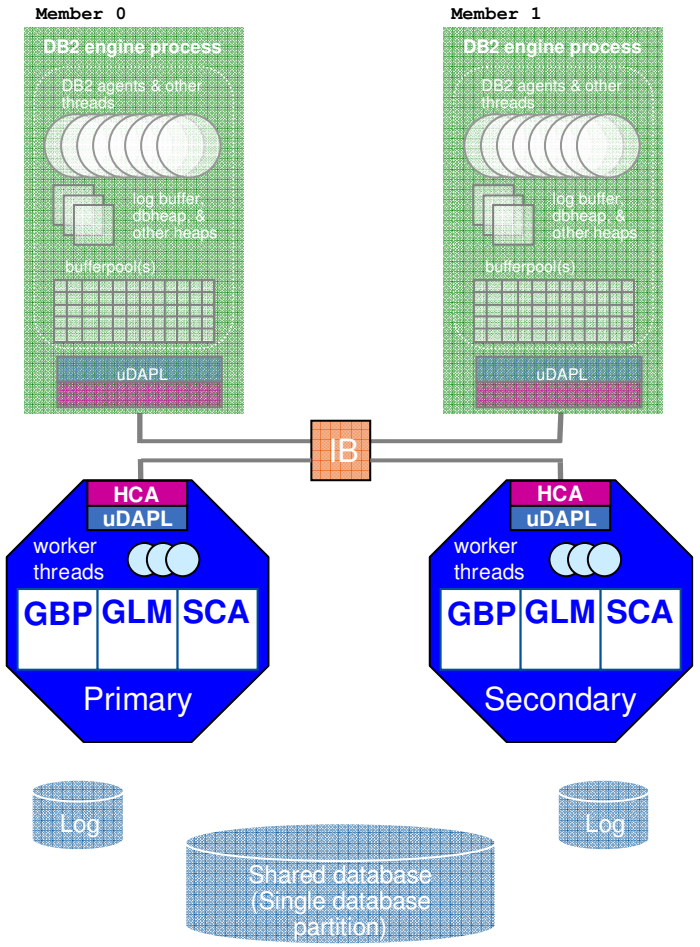
What is a Member ?

- A DB2 engine address space
  - i.e. a DB2 engine process (db2sysc) and its threads
- Each member has it's own …
  - Bufferpools
  - Memory regions
  - Log files
- Members coordinate with each other via the PowerHA pureScale systems, through uDAPL
- Members Share Data
  - All members access the same shared database
  - Aka "*Data Sharing*"



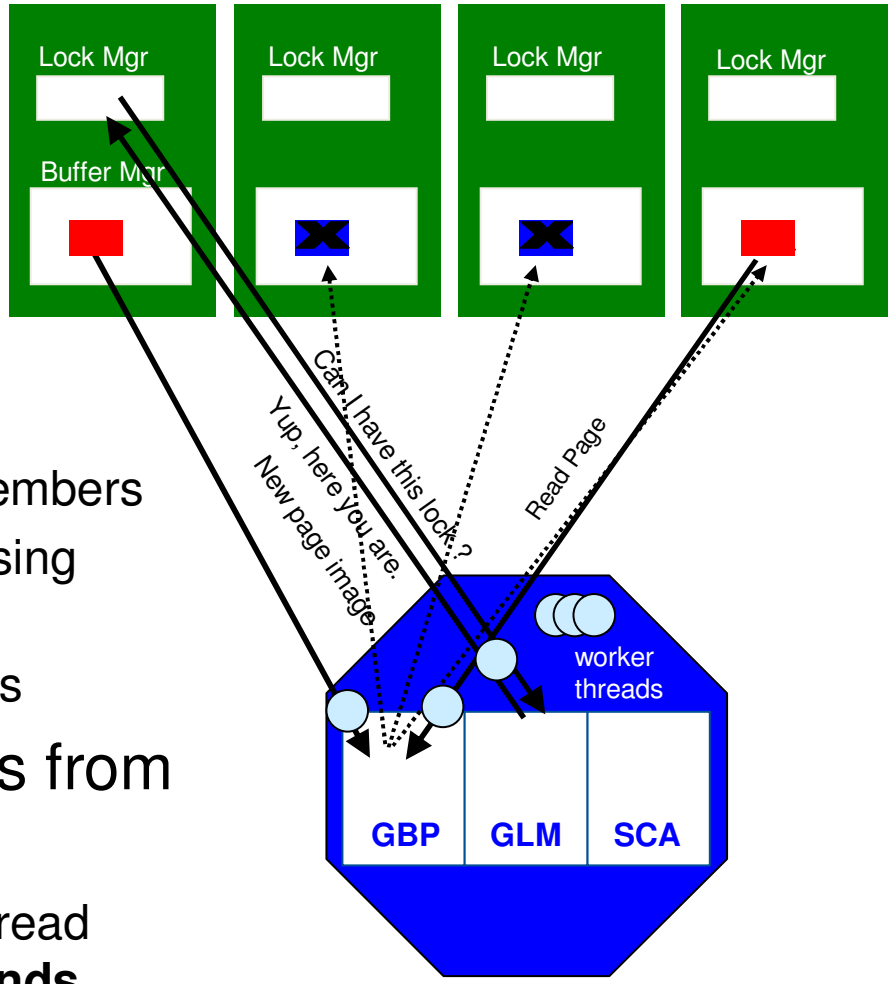
What is a *PowerHA pureScale* ?

- Software technology that assists in global buffer coherency management and global locking
  - Shared lineage with System z Parallel Sysplex
  - Software based
- Services provided include
  - Group Bufferpool (GBP)
  - Global Lock Management (GLM)
  - Shared Communication Area (SCA)
- Members duplex GBP, GLM, SCA state to both a primary and secondary
  - Done synchronously
  - Duplexing is optional (but recommended)
  - Set up automatically, by default



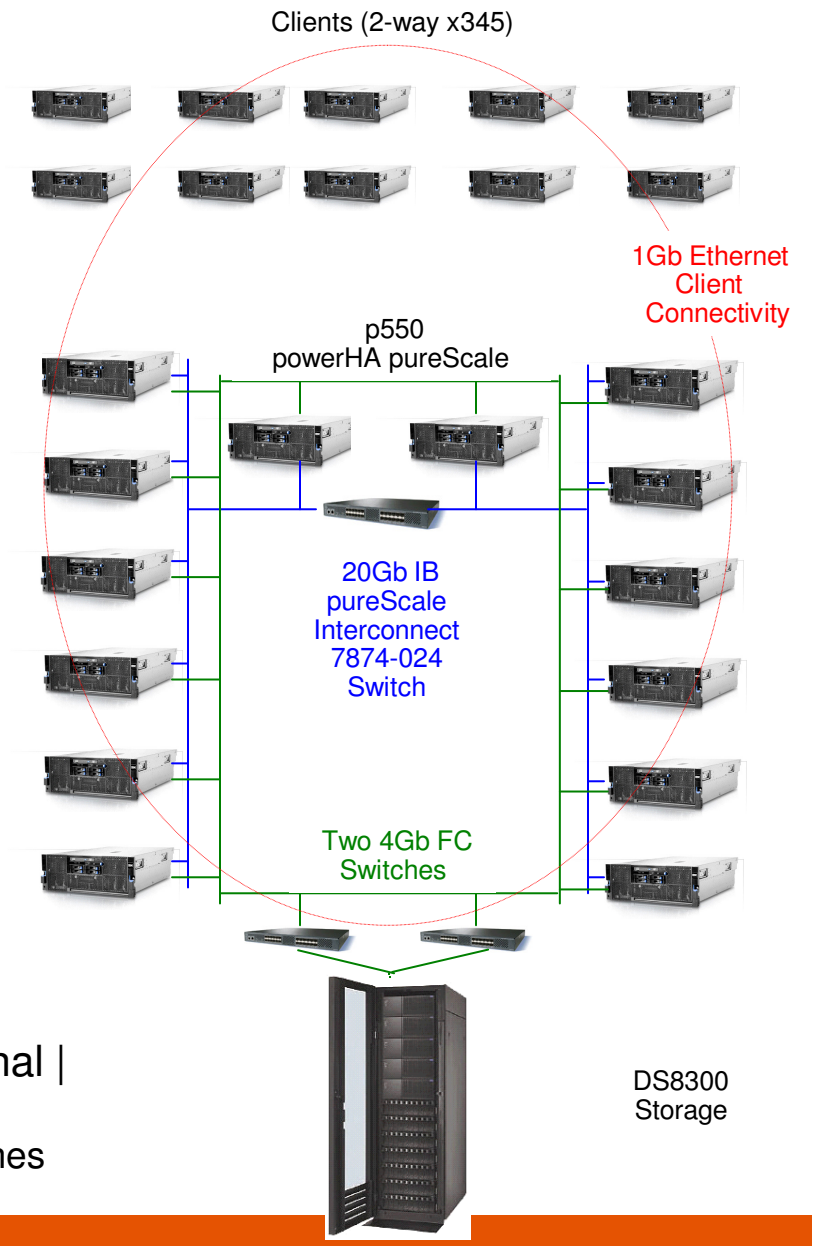
Walking through some key activities

- RDMA exploitation via uDAPL over low latency fabric
  - Enables round-trip response time ~**10-15 microseconds**
- Silent page invalidation
  - Informs members of page updates
  - Requires **no CPU cycles** on those members
  - No interrupt or other message processing required
  - Increasingly important as cluster grows
- Hot pages available to members from GBP memory without disk I/O
  - RDMA and dedicated threads enable read page operations in **10s of microseconds**

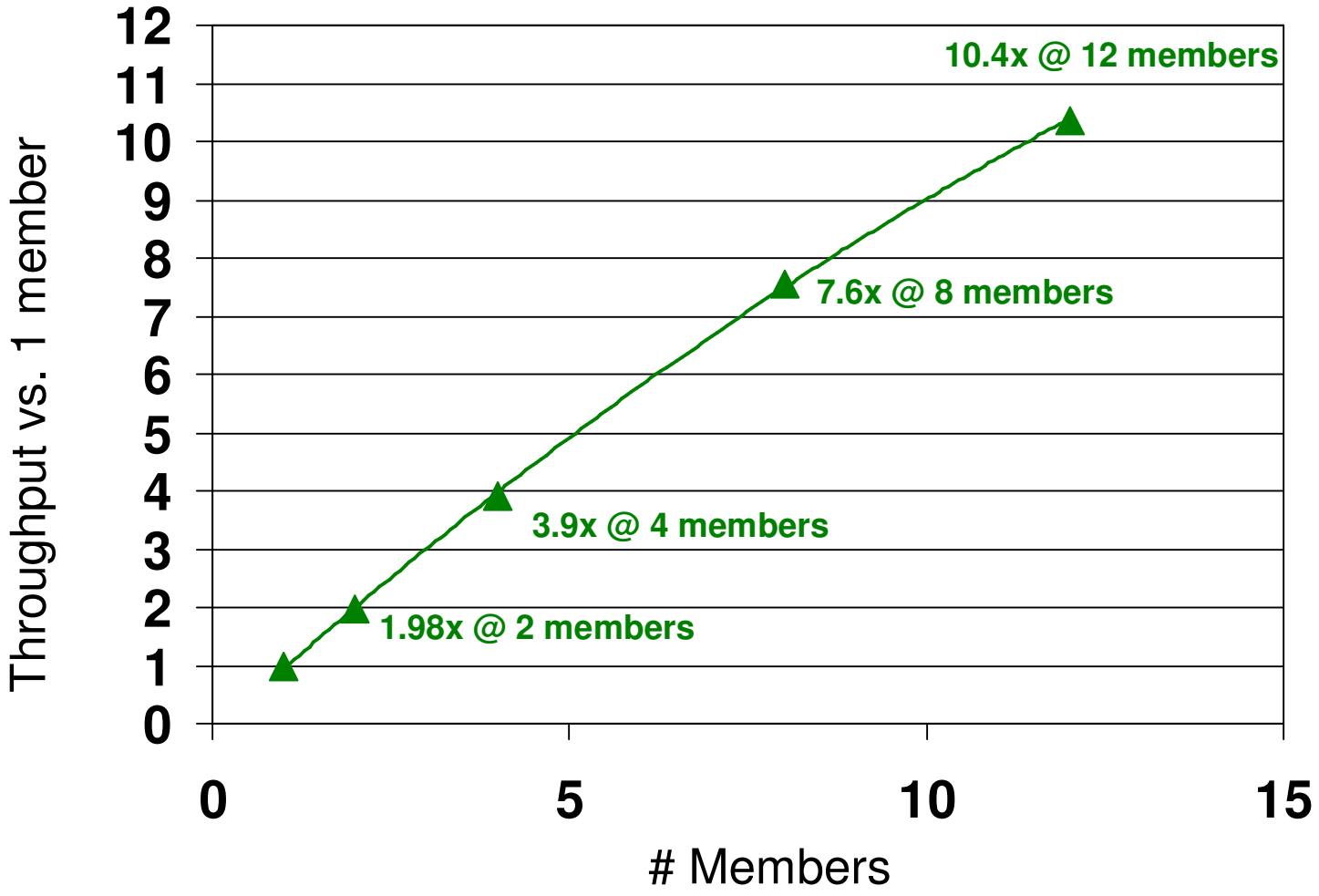


Scalability : Example

- Transaction processing workload modeling warehouse & ordering process
  - Write transactions rate to 20%
  - Typical read/write ratio of many OLTP workloads
- No cluster awareness in the app
  - No affinity
  - No partitioning
  - No routing of transactions to members
- Configuration
  - 12 8-core p550 members, 64 GB, 5 GHz
  - IBM 20Gb/s IB HCAs + 7874-024 IB Switch
  - Duplexed PowerHA pureScale across 2 additional | 8-core p550s, 64 GB, 5 GHz
  - DS8300 storage 576 15K disks, Two 4Gb FC Switches



Scalability : Example

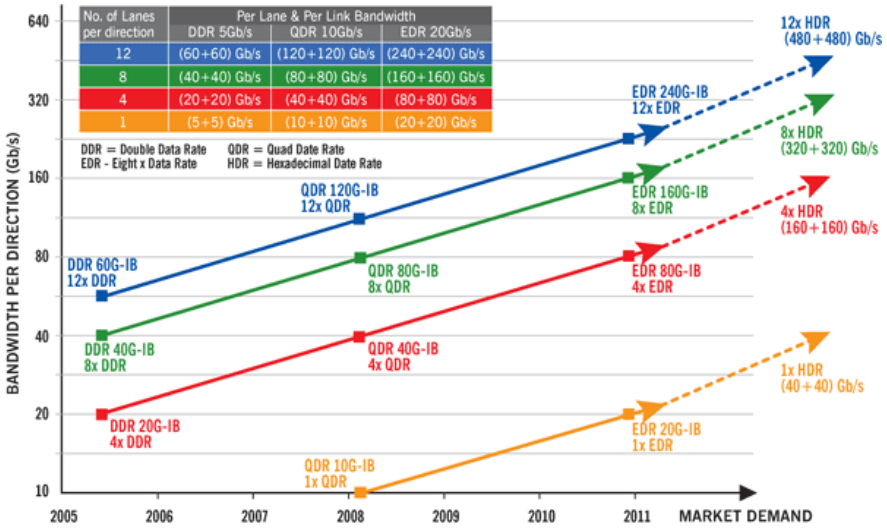


Agenda

- Introduction to pureScale
  - 10,000 ft view
  - Goals & value propositions for customers
  - Technology overview
- Key Concepts & Internals
  - Major components & moving parts
  - Efficient scaling over low-latency interconnect
- → Interconnect issues
  - Requirements, futures, etc.

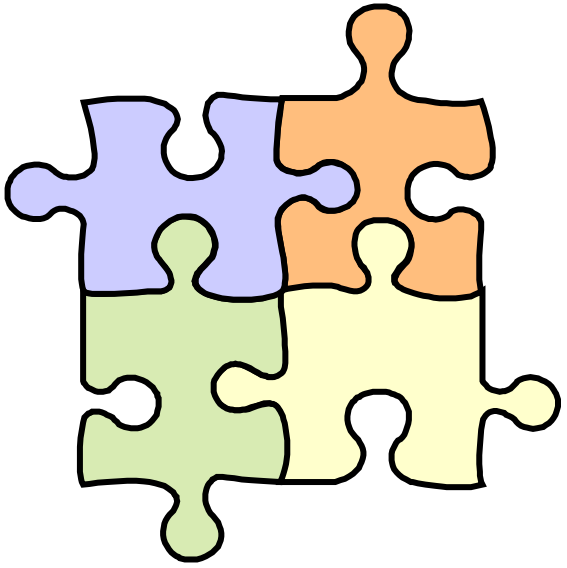
Transport requirements / outlook

- RDMA is a key technology in DB2 pureScale
  - uDAPL + IB spec & maturity make it a natural choice
  - The initial pureScale release platform is well-specified, but fundamentally it is transport-agnostic
- Low latency / high bandwidth is king!
  - QDR / EDR / HDR IB promise improved performance & scalability
  - Stable minimum latencies needed, especially with increased traffic
- Multicast RDMA write



Infiniband Roadmap from www.infinibandta.org

Converged fabrics



- Large-scale database has huge capacity requirement for interconnect, FC disk access, client network connections
- IT policy sometimes frowns on multiple network infrastructures
- Successful convergence needs
  - Sufficient transport capacity – EDR+, 80 Gbps+
  - Fine-grained traffic prioritization
  - Solid QoS mechanisms
- Broad adoption will need demonstrated reliability in heterogeneous, unpredictable IT environments

Link aggregation

- Interconnect fault-tolerance, load balancing
  - Provides needed additional headroom (esp. IOPS) & high availability for the most demanding configurations
  - Standard practice in conventional Ethernet & FC networks
  - Can be handled at application (pureScale) level, but preferable below that

‘Fit and finish’

- 20 years of ‘easy ethernet’ raises expectations
- Fairly high level of expertise required to configure & manage low latency interconnects
  - Very manageable in a lab environment
  - Trickier in a mass-market IT shop
- Stack integration into O/S distributions a great step forward
- Full integration into system management tools still to come

Summary

- A confluence of great technologies
  - High-bandwidth, low-latency interconnect
  - Efficient user-mode transport access via uDAPL
  - RDMA for ultra-lowcost cache coherency
  - DB2 Sysplex

  helps enable pureScale to deliver on its scalability and availability targets
- In addition
  - Maturity / standardization of converged transports
  - Continued IOPS / GbPS growth
  - Improved HA and manageability

  will continue to promote commercial deployment of advanced interconnect stacks