OPENFABRICS ALLIANCE

DB2 pureScale: High Performance with High-Speed Fabrics

Author: Steve Rees

Date: April 5, 2011

Agenda

- Quick DB2 pureScale recap
- DB2 pureScale comes to Linux
- DB2 pureScale and RoCE
- Multi-HCA for increased capacity
- Some futures
- Challenges

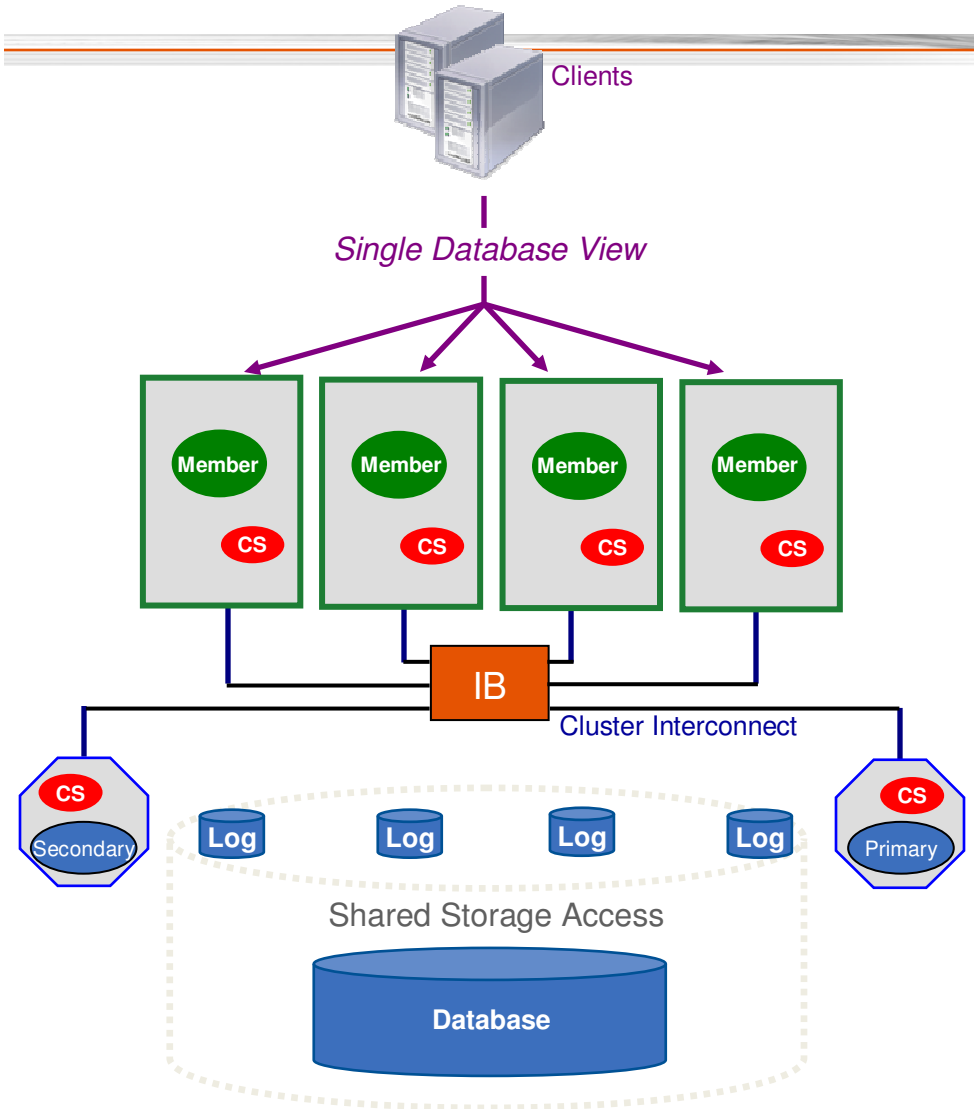
- ***Disclaimer:*** *Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to 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.*

Introducing DB2 pureScale

- Virtually Unlimited Capacity
  - Buy only what you need, add capacity as your needs grow
- Application Transparency
  - Avoid the risk and cost of application changes
- Continuous Availability
  - Deliver uninterrupted access to your data with consistent performance

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. RoCE)

Cluster Caching Facility (CF) from STG

- 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

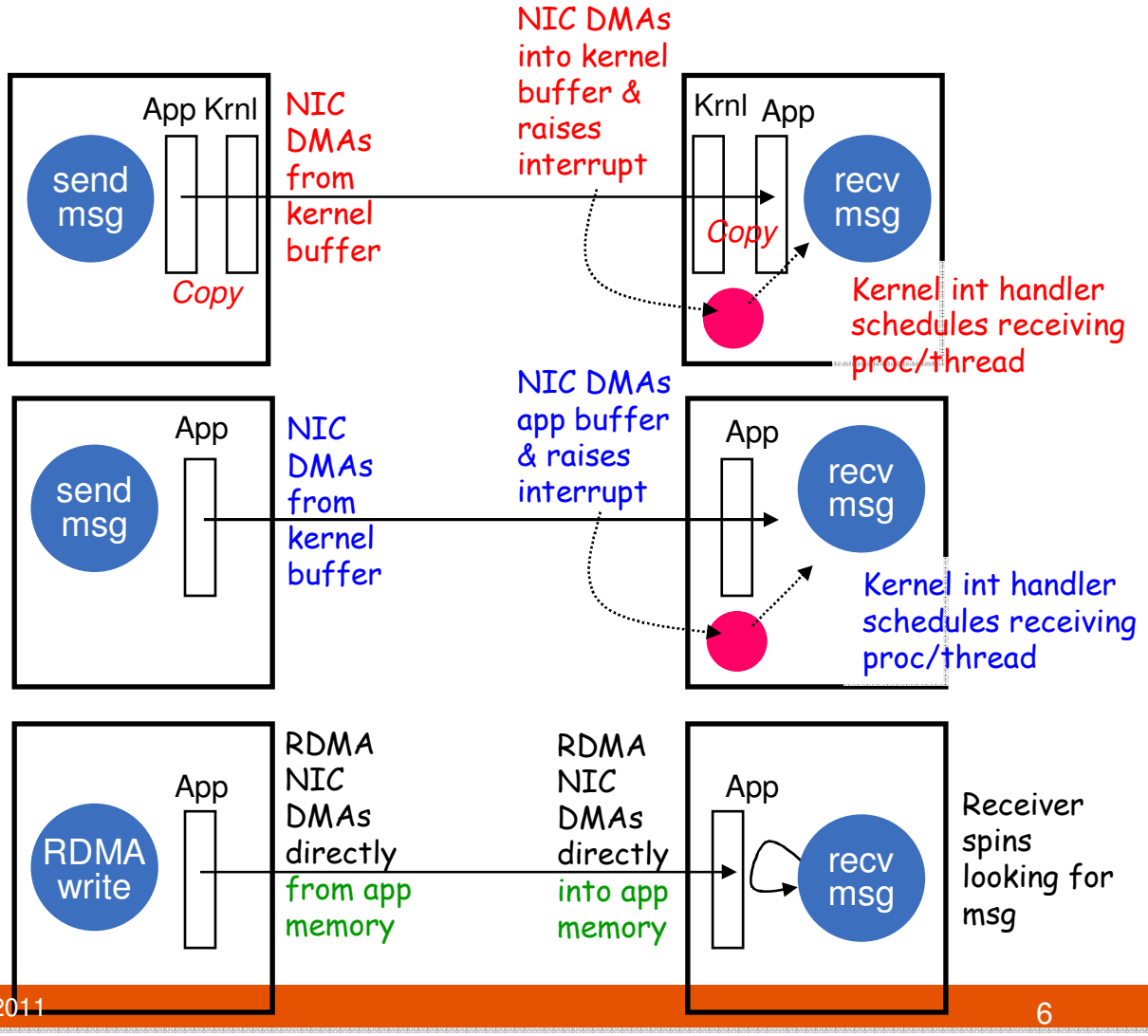
- Failure detection, recovery automation (TSA / RSCT)
- Cluster file system (GPFS)

Sidebar: Send/Receive vs. RDMA

Send / Recv (TCPIP socket)

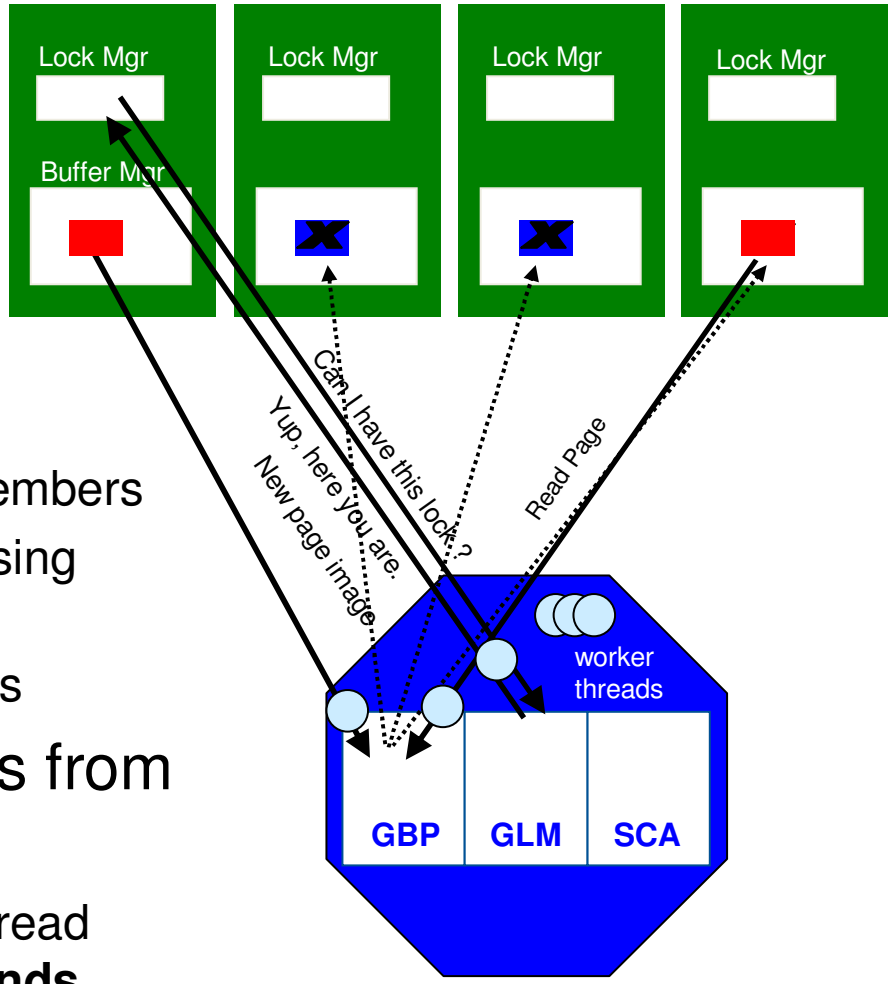
Send / Recv (Reliable Datagram Socket - RDS)

RDMA

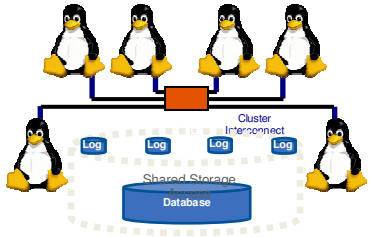


pureScale scales with RDMA & uDAPL

- 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**

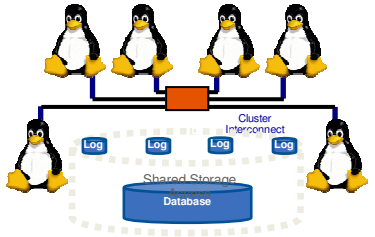


DB2 pureScale on Linux



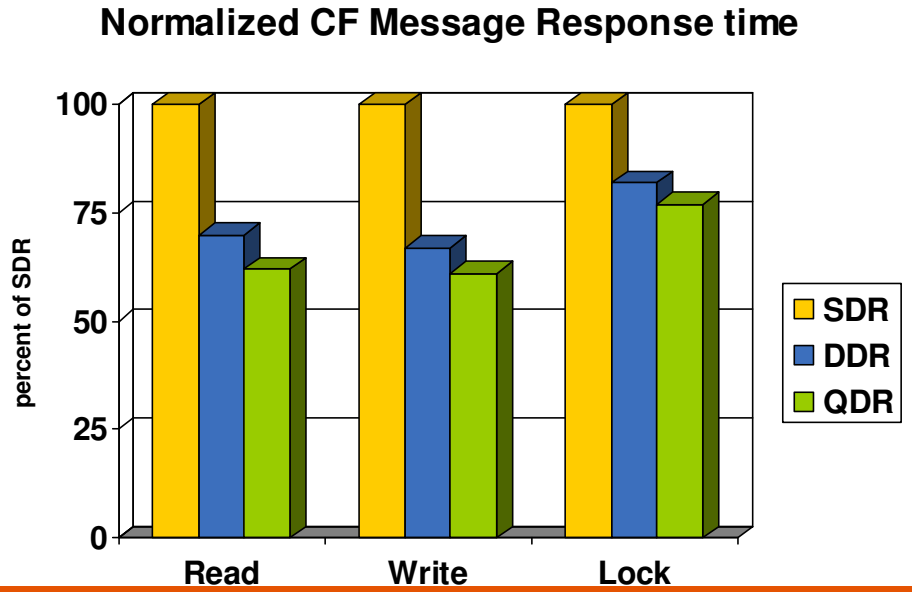
- Customer demand to broaden platform base from initial release on Power AIX
  - Respond to Linux 'sweet spots' for deployment
  - Address customer skill focus areas
- Easy 'port' – in use for years internally for development of pureScale
  - Differences in OFED delivery by distro created some challenges
- Introduced 2010
  - IBM SystemX systems (x3650, x3850, x3690)
  - Mellanox ConnectX-2 QDR IB
  - SLES 10.3 / RHEL 5.5

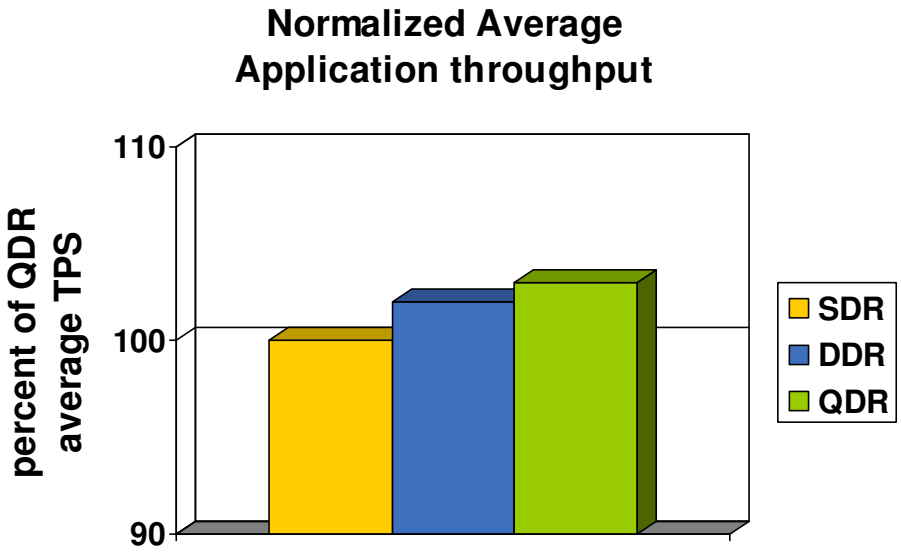
pureScale & Linux & QDR IB



OPENFABRICS ALLIANCE

- Big data movement happens, but small message latency is king
  - Throughput boost in going to QDR – a big win for latency?
  - For *this* workload …
    - Message response time gets a nice boost @ DDR, less @ QDR
    - Overall workload TPS improvement well damped by other factors





pureScale & Infiniband – perfect, right?

- Yes – well, *almost…*

- IB is mature
- IB has obvious technical strengths
- IB is a great fit for a high-performance clustered database

But…

- Some customers are hesitant to deploy a new network type, however wonderful it is

pureScale & RoCE

- ~~IB~~ Ethernet is really mature
- Ethernet is ubiquitous
- ~~IB~~ RDMAoE is a great fit for high-performance clustered database too

- RoCE support added in DB2 pureScale 9.8.0.3
  - Mellanox ConnectX-2 10Gb EN + PFC switch
  - OFED 1.5.2
  - Initially SLES 10

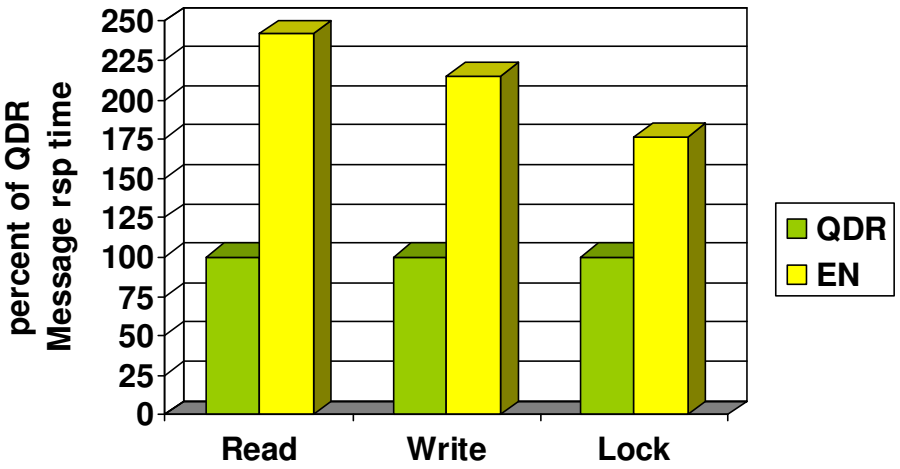
pureScale & RoCE Performance

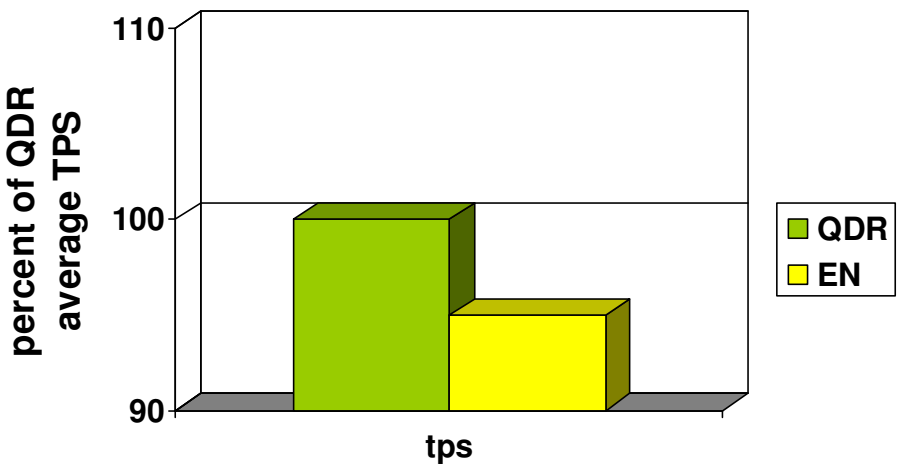
Q: How visible is the difference in nominal bandwidth between 40Gb QDR IB vs 10Gb EN for an average application?

A: Not very … comparable performance to IB makes even 10Gb Ethernet a viable option for many customers

Noticeable at message level

Average tps at application level within 5-10%



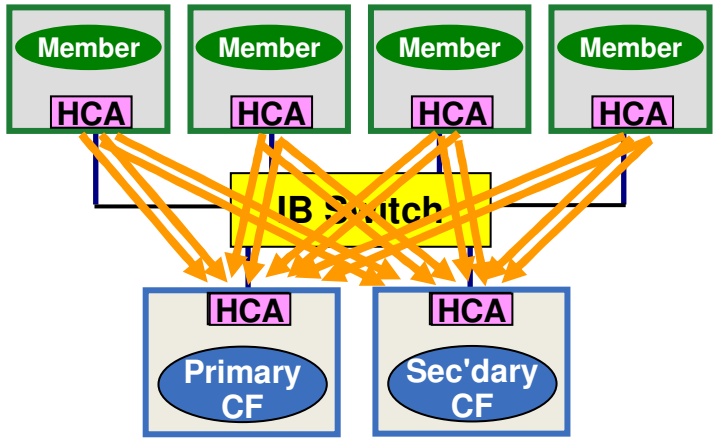


pureScale & Ethernet futures

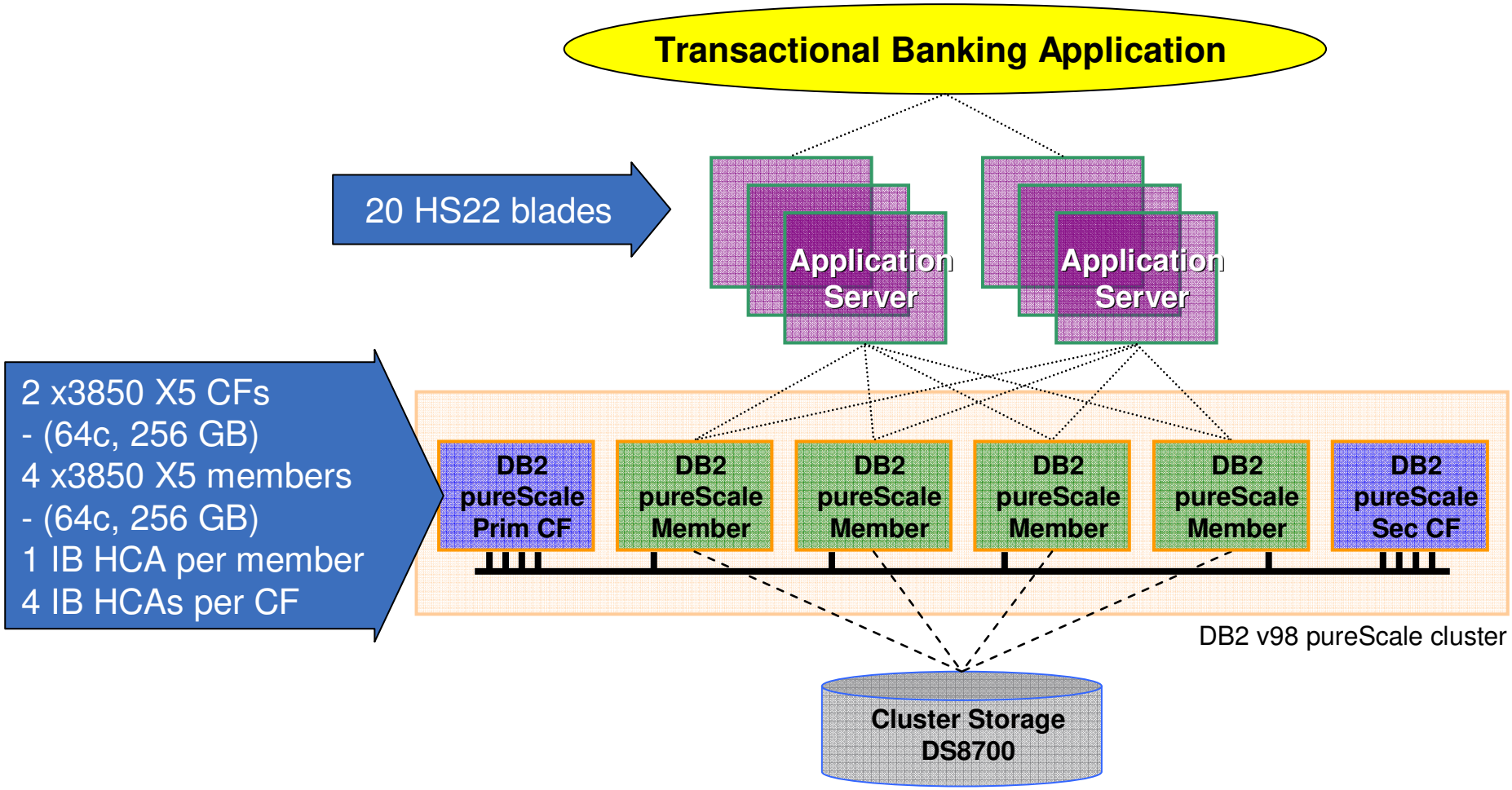
- Strong customer interest encouraging wider support in future pureScale releases
  - Cards, vendors, distros, platforms etc.
- Looking forward to common availability of 40 Gb EN to close gap with QDR IB
  - Larger workloads shipping very large data volumes benefit from the greater throughput
- And what about iWARP?

Multiple CF HCAs

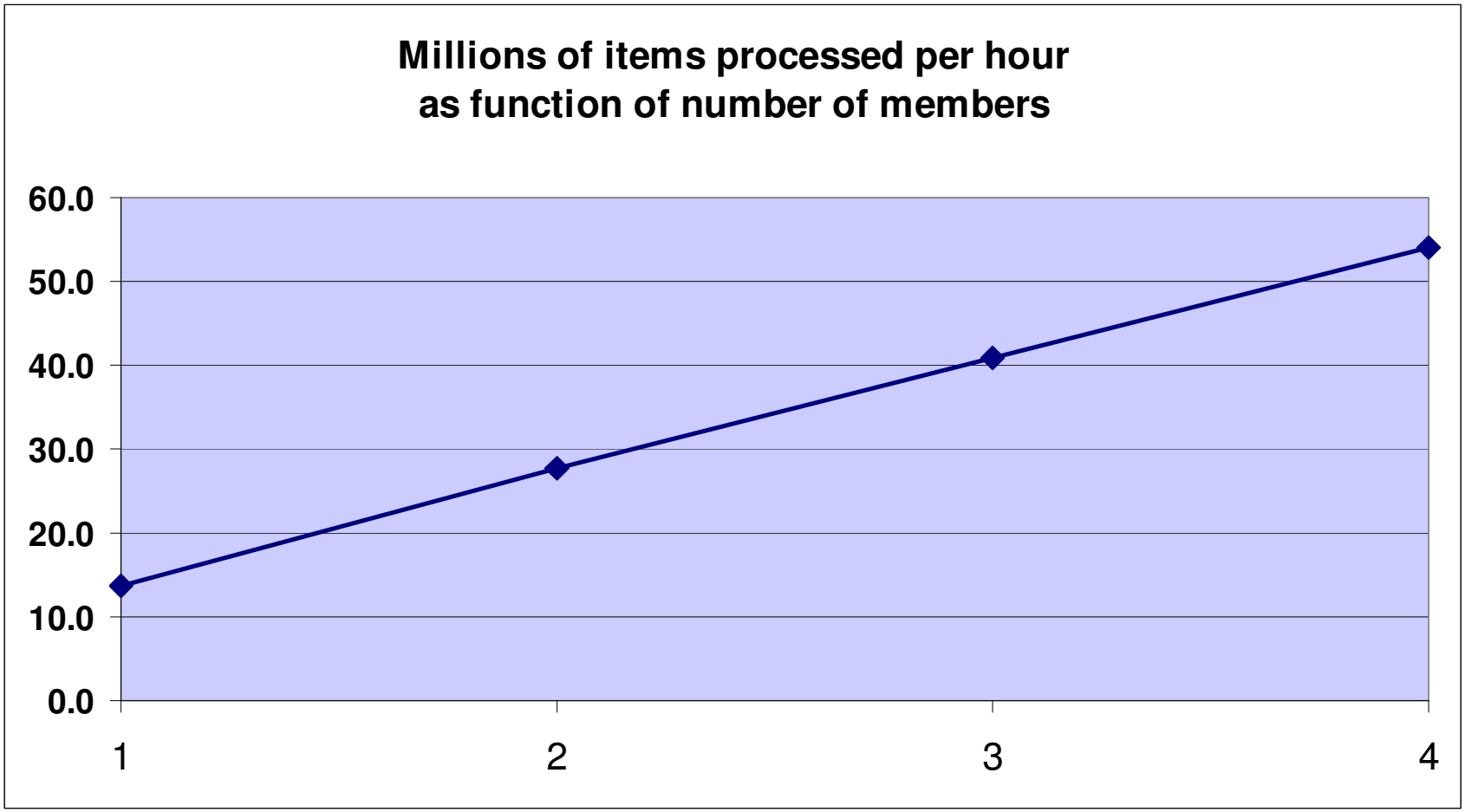
- Low latency to the CF ensures high performance for pureScale
- Duplexed primary & secondary CFs already avoid SPoF
- Very heavy workloads and/or very large clusters could overload the IB / RoCE HCA at the CF
- Multiple CF HCAs in beta fall/2010



Example – pureScale + banking app

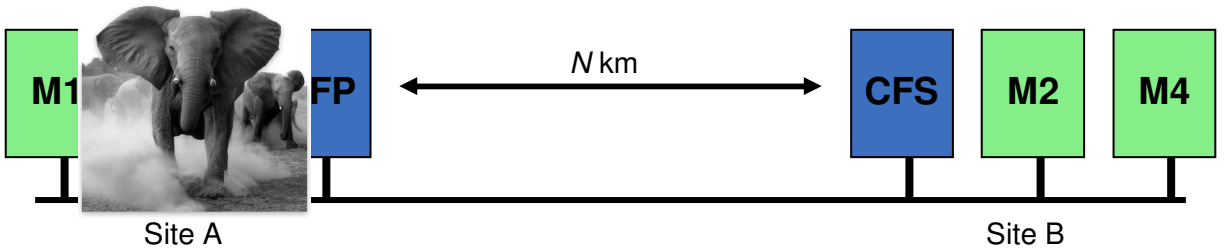


Near Linear Scaling @ 1-4 members

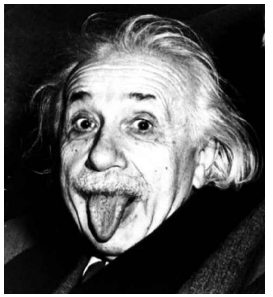


pureScale futures – 'stretched' clusters

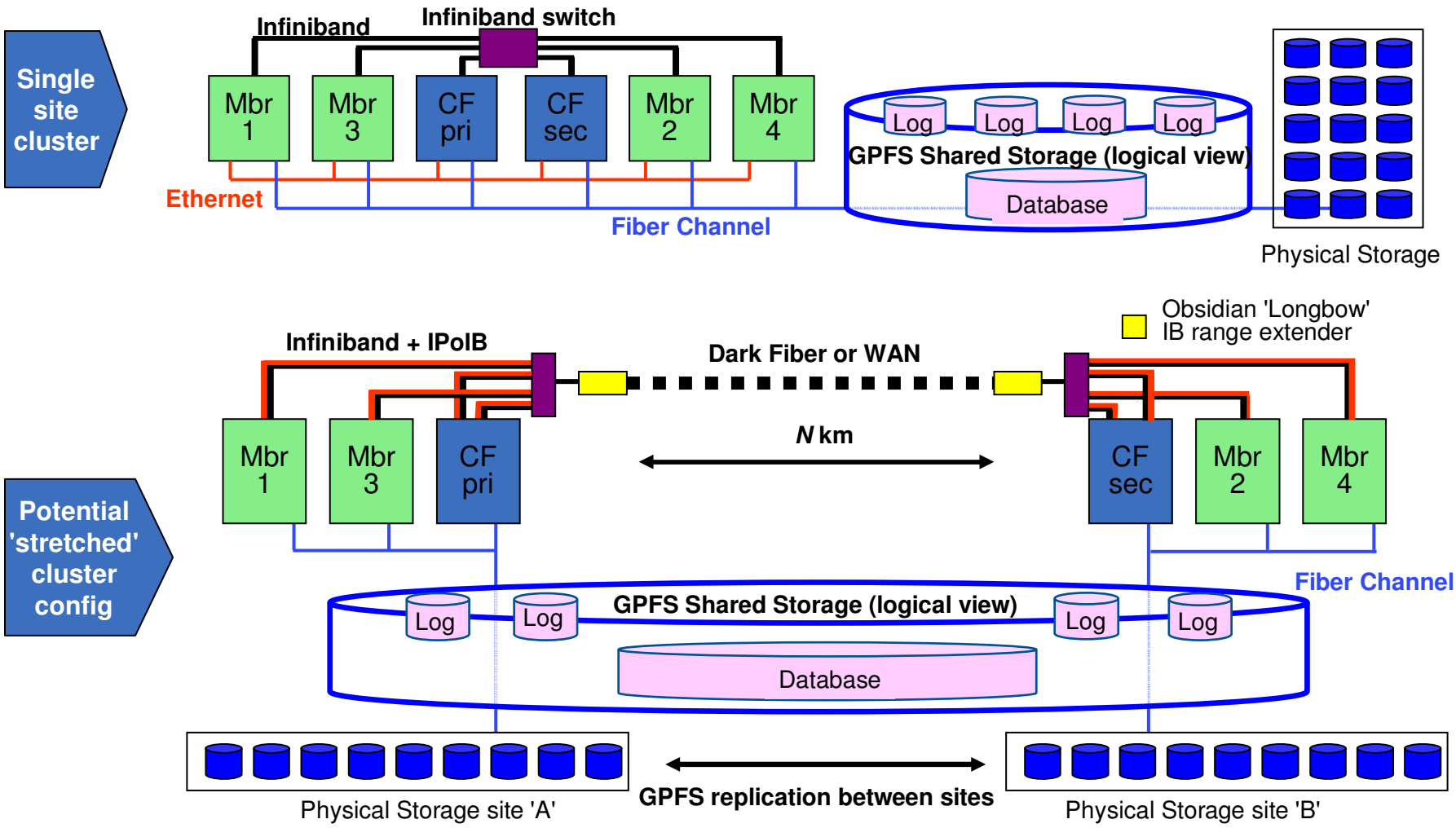
- Splitting the pureScale cluster over two sites offers some disaster resistance
  - Fire, power or communication outage, etc.



- Must be able to stretch RDMA over long distances
  - Currently testing with Obsidian Longbow IB extenders
- Obvious implications from finite speed of light…



Stretching the pureScale cluster



Challenges / observations re: RDMA fabrics

- Inconsistent OFED implementations / packaging across platforms / distros
  - Impediment to porting & commercial DC adoption
- RDMA transports can be challenging to manage
  - Integration with management stacks & basic utilities needed
    - OpenView, Tivoli, even netstat
  - Improving with Ethernet-based implementations
  - Still rough edges around OS & stack integration outside of HPC deployments
- High demand for well-supported virtualization on Linux
  - SR-IOV, KVM, VMware
  - Moving in that direction, but not there yet

Challenges / observations re: RDMA fabrics

- Growth of transport bandwidth Gb/s is goodness, but small message latency is what really counts in many cases
- Adapter bonding required for greater reliability & capacity