November 2003

DB2® Information Management Software

IBM® DB2® for Linux deployed on IBM® BladeCenter Infrastructure

Overview

Boris Bialek
IBM Toronto Lab

1. Introduction

The sizes of databases are growing exponentially every year. For each person on the planet about 150 Megabytes of data area generated per year. A large amount of this data is related to the fast growing use of business applications that enable customers to deliver a foundation for organizational improvements in manufacturing, customer service or data warehousing.

The historical database strategy has been to deploy on large SMP servers with Unix variants such as IBM AIX, HP-UX or Solaris as the operating system. With the tremendous momentum of Linux, more customers are looking to deploy database applications and workloads on Linux. The potential cost savings from initial software licensing, as well as the use of commodity-based servers, is a very compelling value proposition in tough economic times.

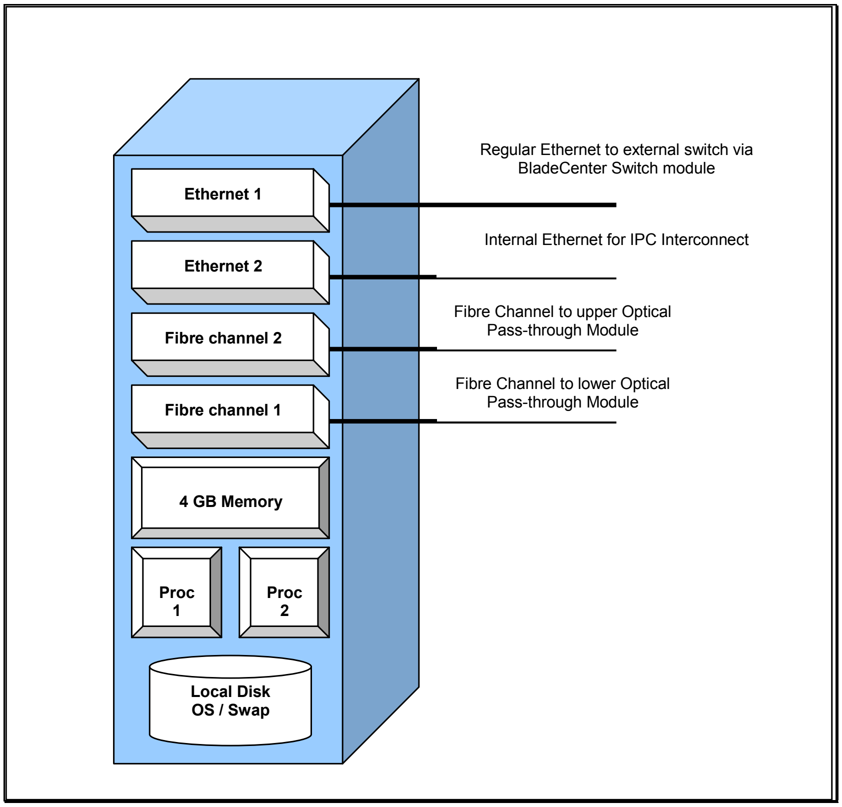
DB2 Enterprise Server Edition for Linux with the Data Partition Feature (DPF) brings the industry leading scalability of DB2 Universal Database, with the ease-of-use and low Total Cost of Ownership (TCO) of Linux into a winning combination for scaleable database deployment. The award-winning DB2 Integrated Cluster Environment delivers a blueprint for a highly scaleable and well integrated Linux database cluster. DB2 ICE allows the flexible use of many different server cluster components, including the IBM BladeCenter®.

Typically, the use of blade servers in the market has been around Web server deployment. In this paper we describe the deployment of DB2 for Linux with the IBM BladeCenter and provide some guidelines for system planning.

2. IBM BladeCenter

The IBM BladeCenter delivers a broad range of helpful features to a database cluster in the areas of manageability and deployment. There are several papers and that broadly discuss these features, and the reader is invited to visit http://www.ibm.com/servers/eserver/bladecenter/ for more information.

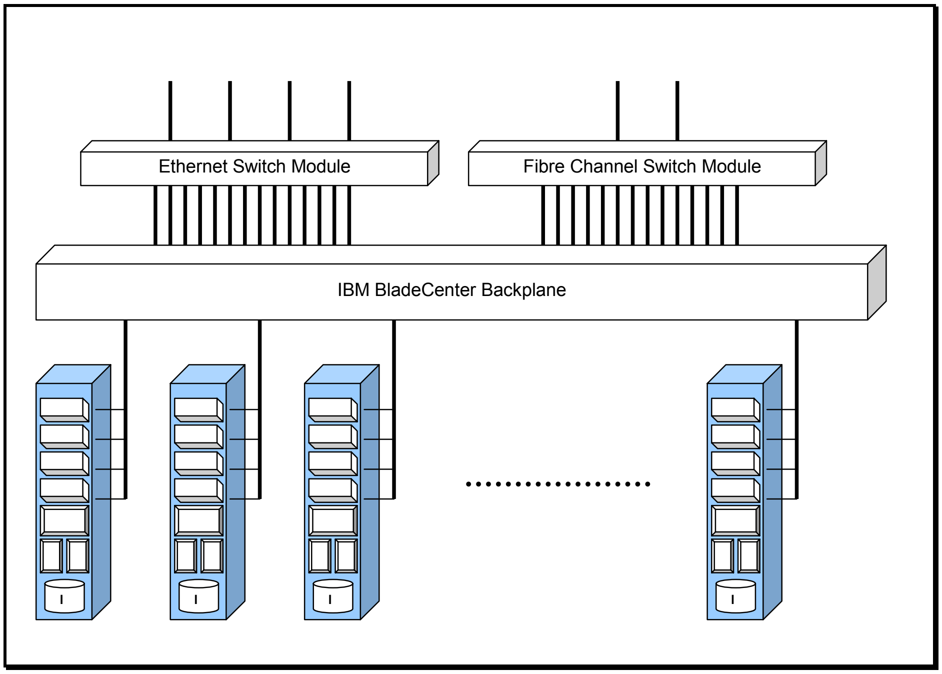
From the database perspective the BladeCenter can be simplified into an assembly of 1U rack-mounted servers as depicted in the following basic diagram:


Ethernet 1
Ethernet 2
Fibre channel 2
Fibre channel 1
4 GB Memory
Proc 1
Proc 2
Local Disk OS / Swap
Regular Ethernet to external switch via BladeCenter Switch module
Internal Ethernet for IPC Interconnect
Fibre Channel to upper Optical Pass-through Module
Fibre Channel to lower Optical Pass-through Module

A standard blade configuration consists of dual processors (e.g. Intel® Xeon processors with 2.4-3.05 GHz clock speed). We know from benchmark experience[1] with DB2 for Linux that each processor is able to utilize roughly 4 GB physical memory per blade. The communication ports onboard are two Ethernet interfaces, equivalent to an IBM xSeries x335 server. The primary port is used for communications between the external application server and the database. The second Ethernet port is used for the inter process communication with DB2. The diagram also shows the use of the Fiber Channel Expansion card, enabling two FCP2 channels for each blade, used for storage communication.

At this point, the picture is identical to the conventional rack-mounted server configuration like the IBM x335. The primary difference is introduced by the components used for connectivity which are integrated into the backplane of the BladeCenter chassis.

[1] Add the reference to the TPC-H benchmark from which the experience referenced is derived.



While the backplane adds the functionality to the BladeCenter configuration, and really is the primary differentiator from typical rack-mounted servers (also know as “pizza box” servers) it introduces design considerations that need to be taken into account when designing an overall database cluster for the BladeCenter. However, with the right selection of interfaces and use of standard expansion cards, these considerations do not limit the deployment of the database at all.

Ethernet

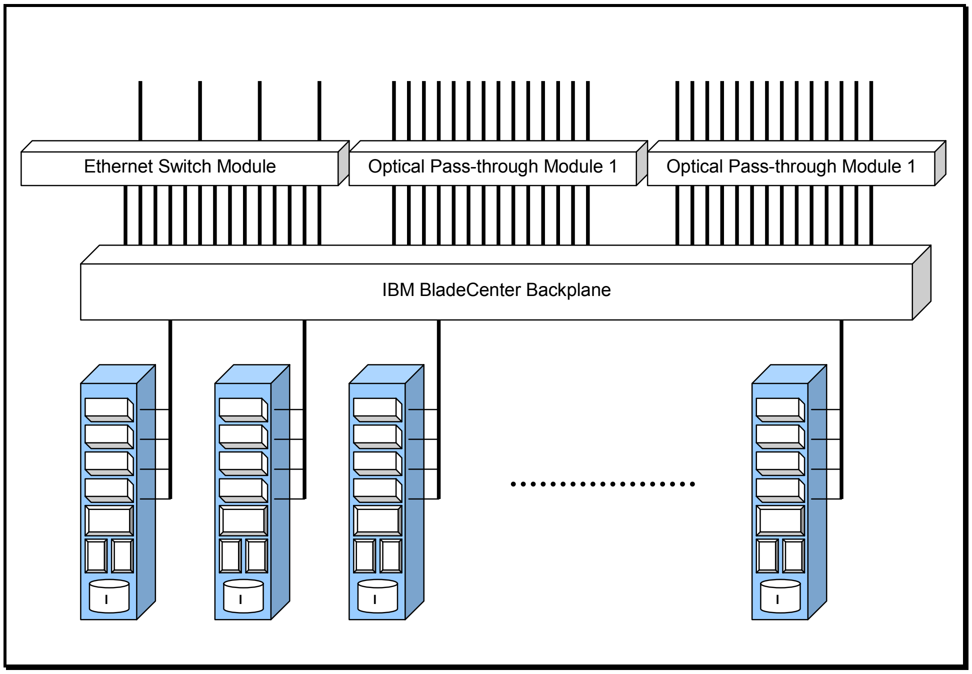
The Ethernet connectivity is implemented with two 1 Gb/s channels from each blade connecting into an Ethernet switch module. The Ethernet switch consolidates the connectivity on the BladeCenter to four ports with 1 Gb/s capacity each, so that the 14 blades share 4 Gb/s bandwidth per each BladeCenter the basic configuration. Additional expansion cards like the Myricom® adapter address this limit but can add considerable expense to the solution, and are not necessarily the choice for every customer. While there should be no actual limit in implementing the application to database server connectivity over Ethernet, the implementation of internal disk storage (ie. iSCSI-based storage) is limited.

The second Ethernet channel available to each blade is ideal for implementing high speed IPC between the blades that is critical for high-performance clustered databases.

Fiber Channel

The Fiber Channel Expansion card adds broader storage IO capacity to the feature set of each blade. These 2 Gb/s FCP2 channels bring the storage IO capacity for the blade back to levels on par with comparable rack-mounted servers. For the outbound connection, two different modules exist.

The first is a Fiber Channel switch module, which is a complete 16-port switch that allows easy connection of other switch-based storage. The disadvantage of the Fiber Channel switch module from the database point of view is the overall bandwidth limitation for each of the nodes. For typical application servers, file and print servers, the 4 Gb/s outbound bandwidth of the Fiber Channel switch is sufficient. The Fiber Channel switch can also be useful for specific deployments of combinations of database and application servers inside a single BladeCenter. This example is discussed in a later section of this paper. However, for high volume databases a better solution is the use of the Optical Pass-through Module (OPM).


Ethernet Switch Module
Optical Pass-through Module 1
Optical Pass-through Module 1
IBM BladeCenter Backplane

The OPM passes through all ports from all Fiber Channel Expansion cards and enables 4 Gb/s IO over the two FCP2 ports per blade. In a conservative sizing, this would allow up to 300 MB/s sustained storage communication per blade, assuming there is enough disk and controller capacity available on the actual physical storage unit to sustain this level of communications.

The illustration above describes a maximum configuration with a total of 14 blades, 56 Gb/s storage communication capacity and sufficient performance for a database cluster with 3 TB raw data[2]. BladeCenters can obviously handle smaller databases as well, but the reader should use this upper limit as a guide in the database planning phase. Multiple BladeCenter units can be connected together to increase the data capacity.

[2] Assuming 28 Fiber channels with 155 MB/s effective throughput and apply normal database sizing rules.

3. Examples of Deployment

Single Database with Multiple Application blades

For many application scenarios a single blade is sufficient to support a larger number of application servers. Experience with SAP R/3 SD benchmarks would set a ratio of 20 application processors to each database processor. Using this ration, one could conceive of a deployment configuration using a two-node database cluster with ten application server blades, and two Web server blades. In this scenario, storage IO performance is not the limiting factor (the typical processor performance and memory can deliver adequate performance in most implementations).

For the application servers the Fiber Channel Expansion card can be eliminated, as the internal spindles are sufficient for loading the application server instance and the Linux operating system. In many cases the application instances are loaded from a single shared NFS volume on one of the blades.

For an application scenario like this the Fiber Channel switch is the preferred option as it allows connectivity into generic storage servers or other systems, and handles automatic failover between the active database blade and failover blades.

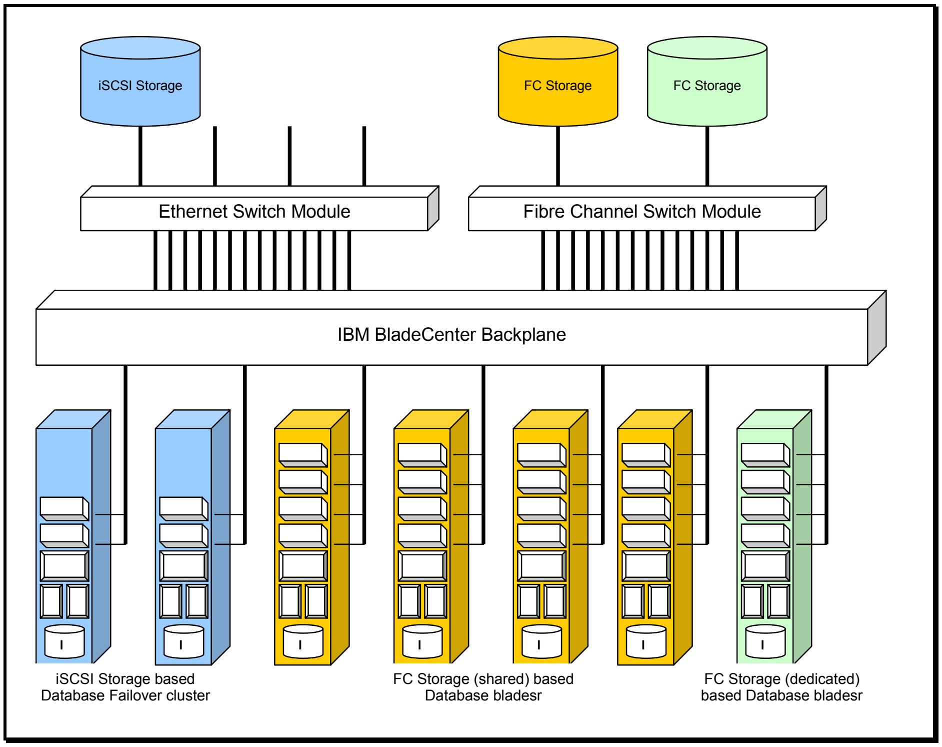
In the case of the OPM, an external switch or switching storage server like the IBM FAStT™ 700 or 900 series is required to ensure failover options for the storage between the two database server blades.

An alternative implementation for this scenario is to use internal iSCSI storage with one of the Ethernet switched ports on the blades. In this case an iSCSI storage server can deliver the necessary performance via Ethernet. This enables the failover of the storage used by the BladeCenter, and the need for any Fiber Channel based components is removed. The downside of an iSCSI implementation is the overhead needed by the TCPIP protocol and Ethernet usage, which reduces the overall available processing power.

Multiple small databases and multiple application blades

An extension to the scenario above would be the implementation of multiple smaller database servers and application servers within a single BladeCenter. In addition to the improved manageability provided by BladeCenter’s extensive integrated capabilities, this allows for rapid deployment of small databases to designated blades for circumstances where a single database system is not possible. Examples would be the need to deploy separate database servers for multiple subsidiaries of the same company, or for multiple customers while leveraging the same infrastructure.

Additional application servers may reside in the same BladeCenter, or additional BladeCenter’s may be deployed to stage the application instances. In this scenario, communication uses the outbound Ethernet. Depending on the database throughput, the implementation of iSCSI or Fiber Channel based storage can be mixed as presented in the diagram below.

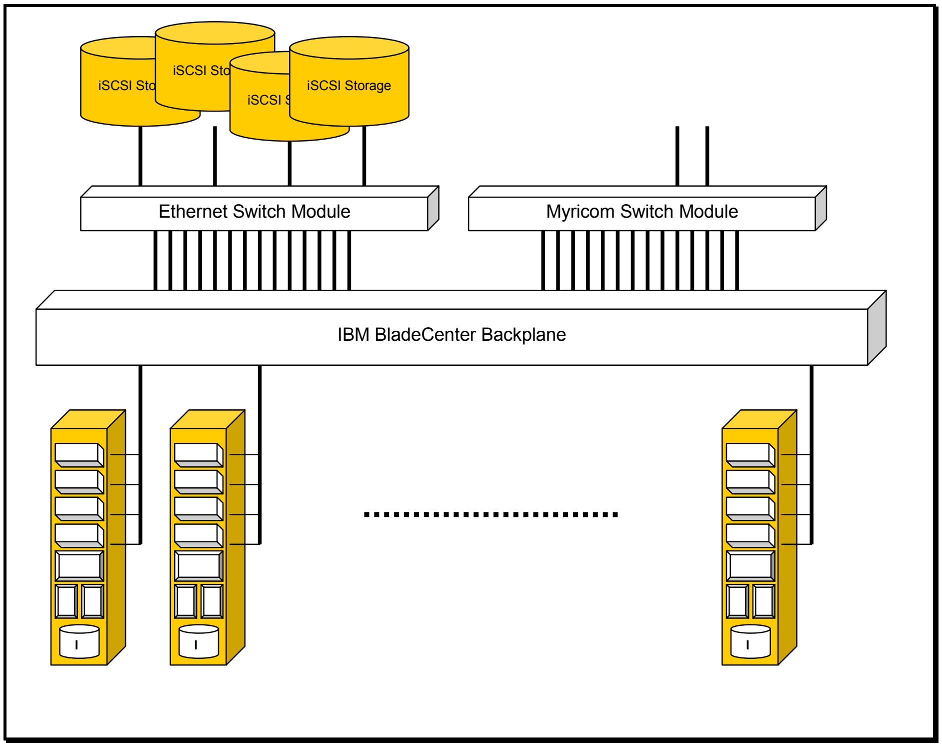

iSCSI Storage
FC Storage
FC Storage
Ethernet Switch Module
Fibre Channel Switch Module
IBM BladeCenter Backplane
iSCSI Storage based Database Failover cluster
FC Storage (shared) based Database bladesr
FC Storage (dedicated) based Database bladesr

In this illustration, a two blade failover cluster (blue) is connected to a single iSCSI channel. Four blade servers (orange) share one of the FC connections and one blade server (green) has a dedicated connection to a FC storage system. The overall storage connectivity in this case is 5 Gb/s over seven servers which is sufficient for average load, with some peak processing requirements directed to some of the databases. We would need the Fiber Channel Expansion card for five of the servers, and the FC switching module is still needed.

For a more consolidated approach the implementation of the Ethernet Expansion card would be a second option. Additionally this would allow standardizing on the iSCSI storage protocol which has a lower price point, but would have the same amount of IO capacity with four ports at 1 Gb/s capacity each (equal to the two FCP2 channels from the FC switch).

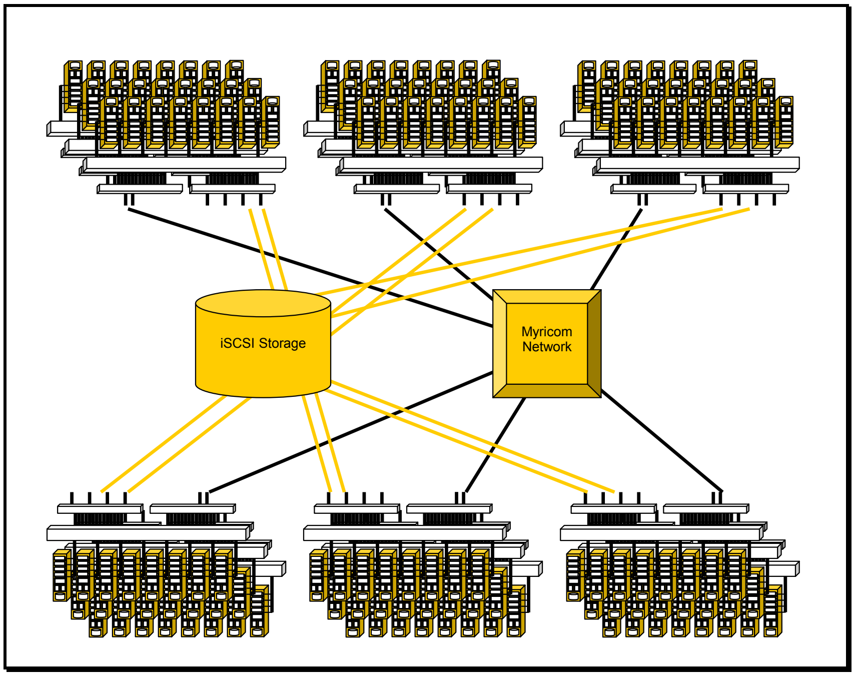
Compute-intensive database cluster

The next level of cluster is the deployment of DB2 for Linux across a number of blades with the focus on compute-intensive applications. The typical application representing this configuration would be CPU intensive operations on the local data set residing on each single blade, while allowing complete access to the full data set across the BladeCenter. An example of this configuration is the implementation of DB2 for Linux clusters as part of a Linux Beowulf high performance computing environment. This can be a wind tunnel or a seismic exploration application. The data used by these applications were historically stored in flat files. With the introduction of advanced mining capabilities in commercially available relational databases, they are now are being used to enable advanced function (like cross-correlating multiple data sets within a single database).



A DB2 for Linux cluster solution, such as DB2 ICE, delivers the data repository for the application framework. Its storage IO configuration is lower than the requirements for more data intensive applications in the business segment. One expansion card slot is needed for the Myricom HPC Expansion card. This leaves only the local hard drive or iSCSI storage for storage IO. The bandwidth and latency of the Myricom network allows for the distribution of DB2 data across a large number of multiple BladeCenters as shown in the diagram below.

Existing Beowulf configurations can reach up to several thousand nodes in one single cluster (see Top 500 high performance computing list under www.top500.org for details) and DB2 ICE for Linux matches this same concept you see in a complete data grid.

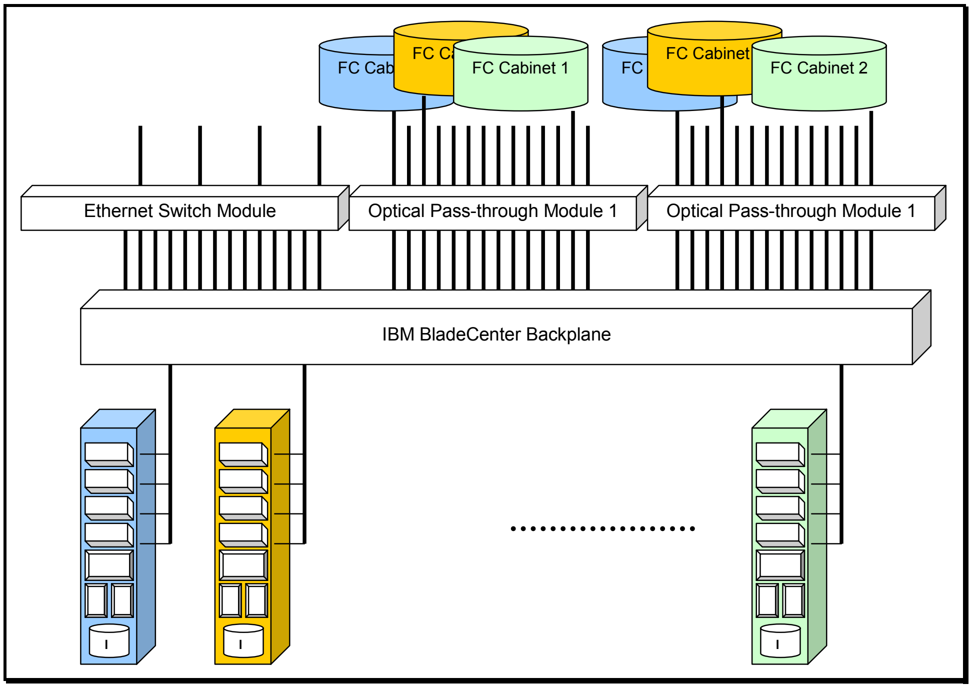


Typical extensions of DB2 would include DB2 DiscoveryLink, Information Integrator or the DB2 Netsearch Extender.

Data Warehousing

While data warehouses in the tens of terabyte are more commonplace today, the more common data volume is less than 3 TB. The IBM BladeCenter allows a best of breed implementation for those warehouses with the smallest physical footprint possible. The excellent price/performance makes it a perfect migration platform for NCR Teradata and other dated technologies from vendors who aren't making the investments in base server infrastructure.

DB2 ICE for Linux deployment uses all the same planning criteria discussed earlier for typical "rack-mounted" server deployment.


FC Cab
FC Ca
FC Cabinet 1
FC
FC Cabinet
FC Cabinet 2
Ethernet Switch Module
Optical Pass-through Module 1
Optical Pass-through Module 1
IBM BladeCenter Backplane

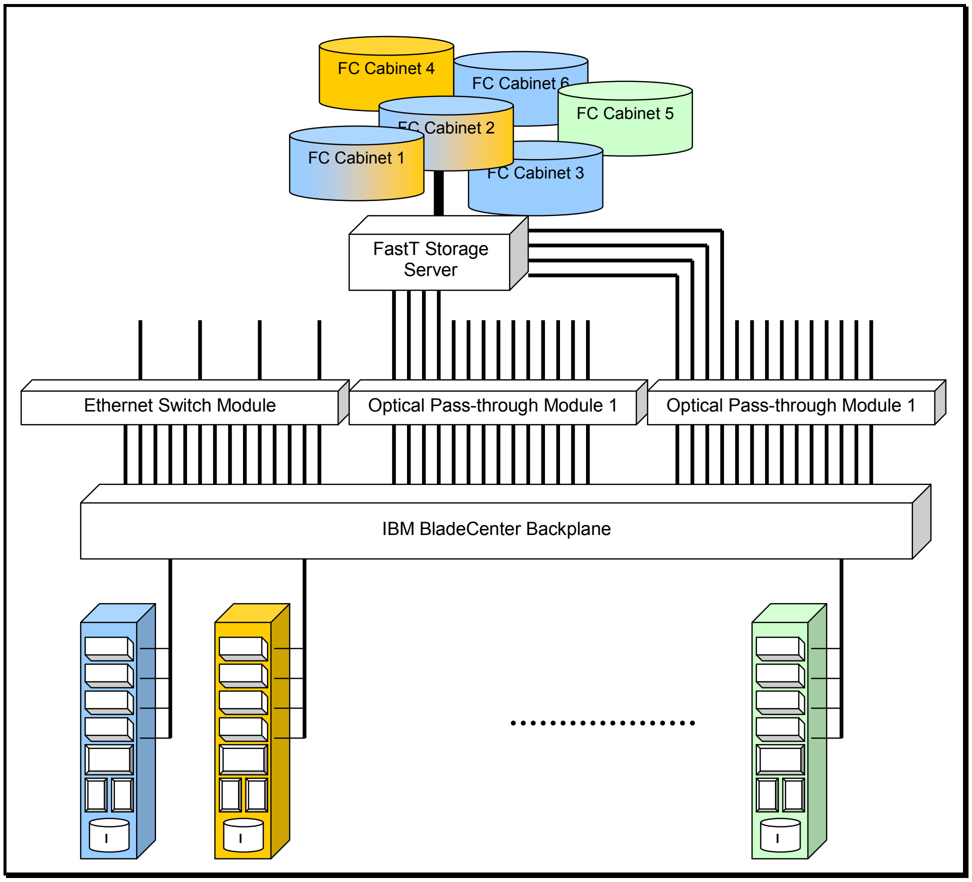
The best price/performance is delivered with direct attached Fiber Channel storage in expansion cabinets without any additional controller. The Linux operating system manages the disks directly as single spindles through the Linux volume manager (LVM) that comes standard with any Linux distribution. The LVM has been proven in various benchmarks in RAID1 configurations as a solid performer, as well as a reliable solution[3]. This configuration is not focused on high availability or flexibility in the management of the available storage. Customers implement this type of configuration for environments where 24x7 availability is not the critical requirement, but rather ease of implementation, expandability through the adding of additional nodes, and low system costs are the primary decision criteria.

High availability can be achieved through implementing redundant paths to the disk cabinets. Each blade manages only one single disk cabinet and uses the second FC port to connect to a disk cabinet from a second blade. While this reduces the available storage IO capacity per blade to a maximum of 2 Gb/s (1/2 of the standard 4 Gb/s), it facilitates use of simple Open Source failover software while providing high availability for the database cluster.

The most sophisticated configuration for the IBM BladeCenter with DB2 ICE for Linux uses intelligent storage servers or SAN's. In this case the storage IO capacity of 4 Gb/s can be sustained in high availability environments. In addition, we gain flexible storage management for the available subsystems.

[3] for example http://www.tpc.org/tpch/results/tpch_result_detail.asp?id=103072902)

The solution with a direct attached intelligent storage server, like an IBM FAStT storage server, enables the highest level of availability, including redundant mirroring to remote sites for disaster recovery. Furthermore, it allows a very flexible distribution of the available storage to the various blades while having sustained storage IO capacity.



The configuration presented in the diagram enables failover between the two OPMs, redundant paths to the disk cabinets and flexible allocation of the actual disks with the blades. Taking things to the highest level of redundancy, one would deploy a redundant storage system with a redundant BladeCenter for disaster recovery.

4. Conclusion

The IBM BladeCenter and IBM DB2 Integrated Cluster Environment provide an elegant, low-cost solution for database clusters in the 3 TB range. It allows for a flexible deployment of a broad range of applications and scenarios ranging from the small scale SAP implementation, to midrange data ware housing in the multi-Terabyte class.

The combination of the IBM BladeCenter and IBM DB2 ICE for Linux is a combination of award winning hardware and database software, delivered by a technology partner you can trust for support of the Linux environment.

The choice of applications is not limited to the examples in this paper, but the considerations discussed should provide the reader a solid baseline for planning and deployment.

IBM DB2 and xSeries technical specialists are happy to assist with more detailed planning and implementation.

.

IBM Canada
8200 Warden Avenue
Markham, ON
L6G 1C7
Canada

Printed in United States of America
10-03

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates.
The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.
Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

The information in this white paper is provided AS IS without warranty. Such information was obtained from publicly available sources, is current as of 01/30/2003, and is subject to change. Any performance data included in the paper was obtained in the specific operating environment and is provided as an illustration. Performance in other operating environments may vary. More specific information about the capabilities of products described should be obtained from the suppliers of those products.