Commercial Linux Clusters

A guide for everyday e-business clusters

Internal use, but not confidential. Last update 2006-03-07.

Avi Alkalay <avix@br.ibm.com>
Linux, Open Standards Consultant
IBM Corporation

About this Document

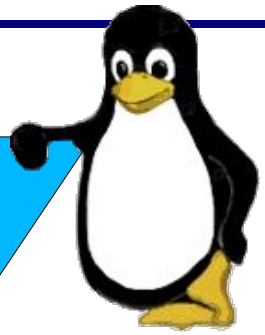
- This document highlights **High Availability Linux Clusters** on IBM hardware and software, covering different cluster solutions
- It is intended to be used by architects, sales people and customers
- Last version available at: *http://avi.alkalay.net/linux/docs/ha/*
- The e-server logo uses a special true-type font that can be installed from *http://watgsa.ibm.com/~avibrz/public/c4eb/ibmfonts*
- For Linux viewers, you should install Microsoft fonts for better results from *http://avi.alkalay.net/software/webcore-fonts*

High Availability Status

- **Who wants Low-Availability systems?**

- **Why are there so few High-Availability systems?**

Linux High Availability Potential



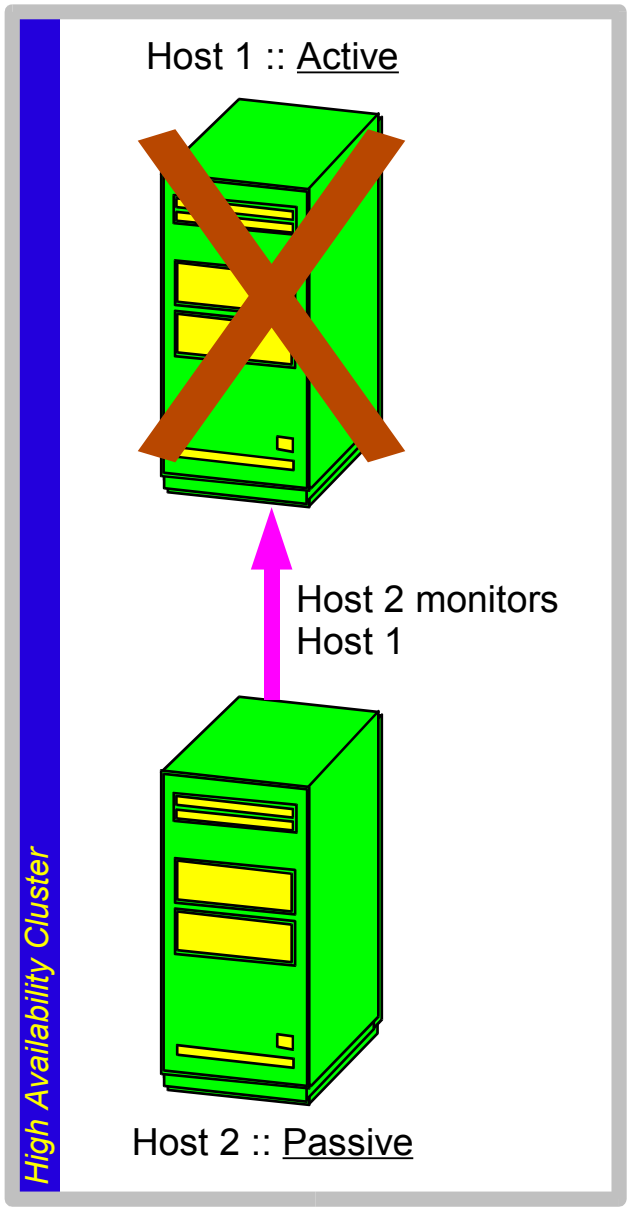
HA on Linux

- ✔ Unexpensive Hardware
- ✔ Free HA Software
- ✔ Simple
- ✔ All systems

Old HA

- ✗ Expensive Hardware
- ✗ Expensive HA Software
- ✗ Complex
- ✗ Exclusive systems

How High Availability Works ?



1. Host 1 is the main application server. HA Software on Host 2 monitors Host 1 availability
2. When Host 1 fails, the HA software automatically takes the following actions
3. HA Software makes the *Application* Data Files available on Host 2
4. HA Software sets *Application* service IP addresses on Host 2
5. HA Software starts *Application* on Host 2

Run in Screen Show Mode

Which Applications I Can Cluster ?

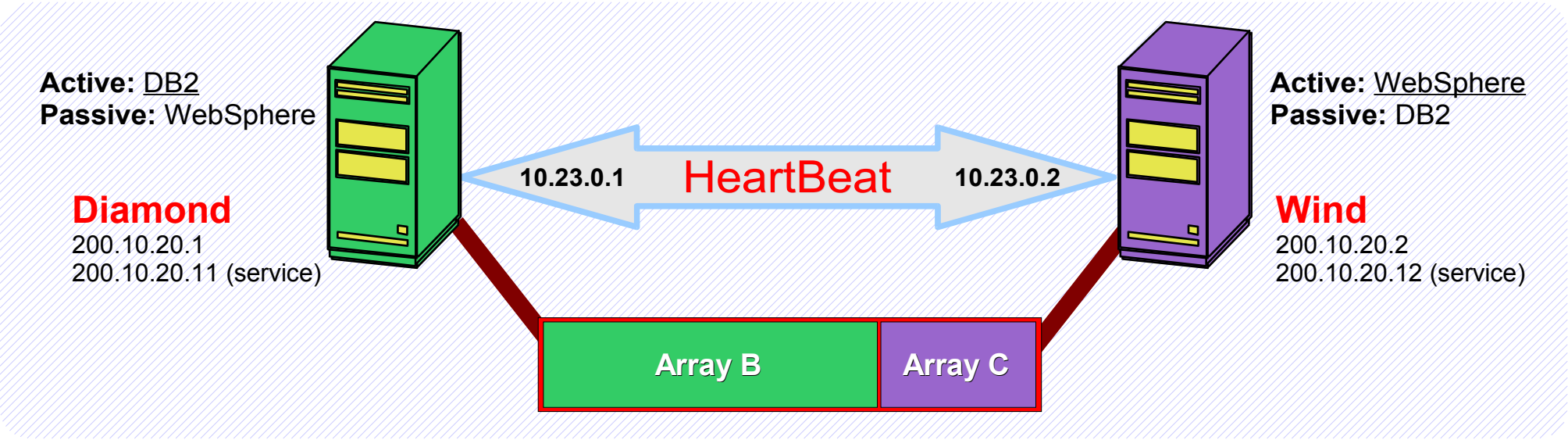
- Any Server Software can be **HA** clustered
  - Any DB, Web-Server, WebSphere, Domino, Samba or other File Server, Mail Server, ERP, etc. Application doesn't have to be aware of its HA cluster context
- If some server software can't be included in an HA context, the client protocol with the server must be reviewed. This is very rare.
- An application (and filesystem) must be able to recover from crashes – preferably quickly. Any regular DB executes this action when started.
- All nodes may run different active applications simultaneously. The peer node will be passive for this particular app. These are Active-Active clusters

A Note About Parallel Applications

- There is some special apps that were built to use other nodes' computing capacity in a parallel way
  - Example: DB2 Parallel Server, Oracle RAC, Domino replicas, etc
- These apps use to have a Controlling Process that runs on a unique node and dispatches activity to slave nodes
- High-Availability must be provided for the Controlling Process in the same way as for regular apps
- Parallel Applications use to require (but not always) special storage configurations, for simultaneous multinode shared data access

HA Common Architectures

Cluster Architecture Perspectives



Service

DB2 Service
- Initially active on Host Diamond
- **Service hostname:** db2.domain.com
- **Service IP:** 200.10.20.11

WebSphere Service
- Initially active on Host Wind
- **Service hostname:** was.domain.com
- **Service IP:** 200.10.20.12

Host Oriented

diamond.domain.com
- **IPs:** 200.10.20.11 (service), 200.10.20.1
- **HB:** 10.23.0.1
- **Initial service:** DB2

wind.domain.com
- **IPs:** 200.10.20.12 (service), 200.10.20.2
- **HB:** 10.23.0.2
- **Initial service:** WebSphere

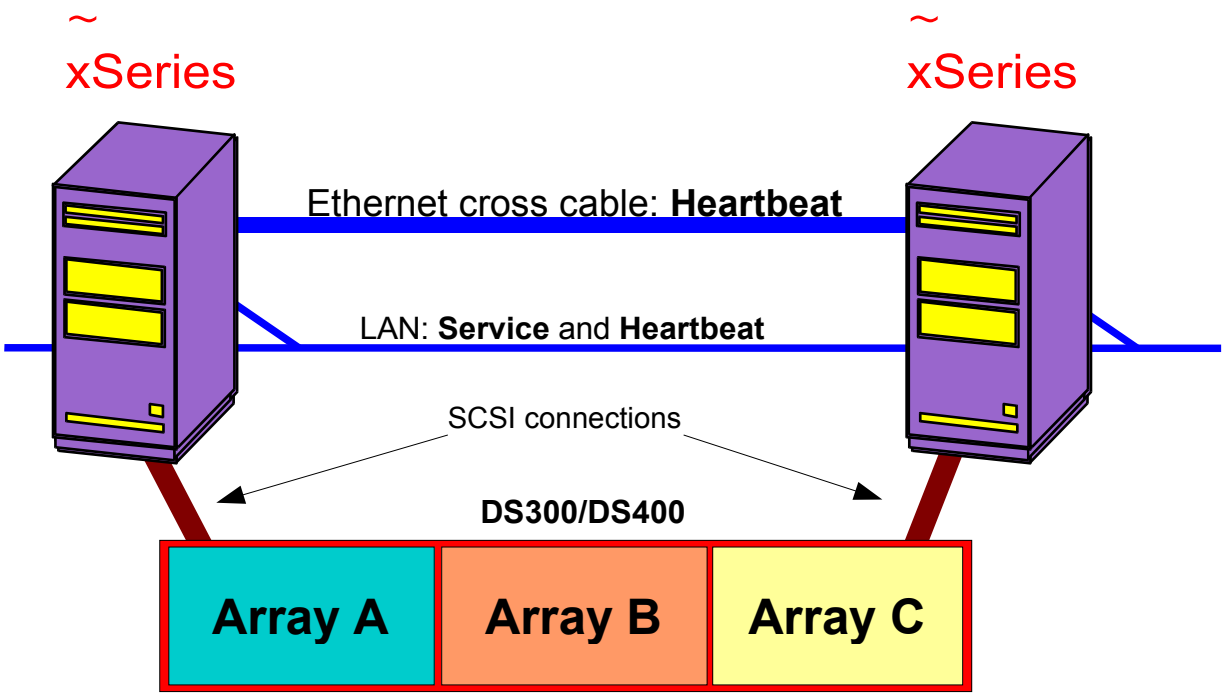
Array Oriented

Array A :: Linux OS, Application Software
- Internal storage. Each machine has its own.

Array B :: Database
- Online for Diamond. Offline for Wind (on if Diamond fails)

Array C :: Application business logic files
- Online for Wind. Offline for Diamond (on if Wind fails)

ServeRAID and SCSI Clusters Overview



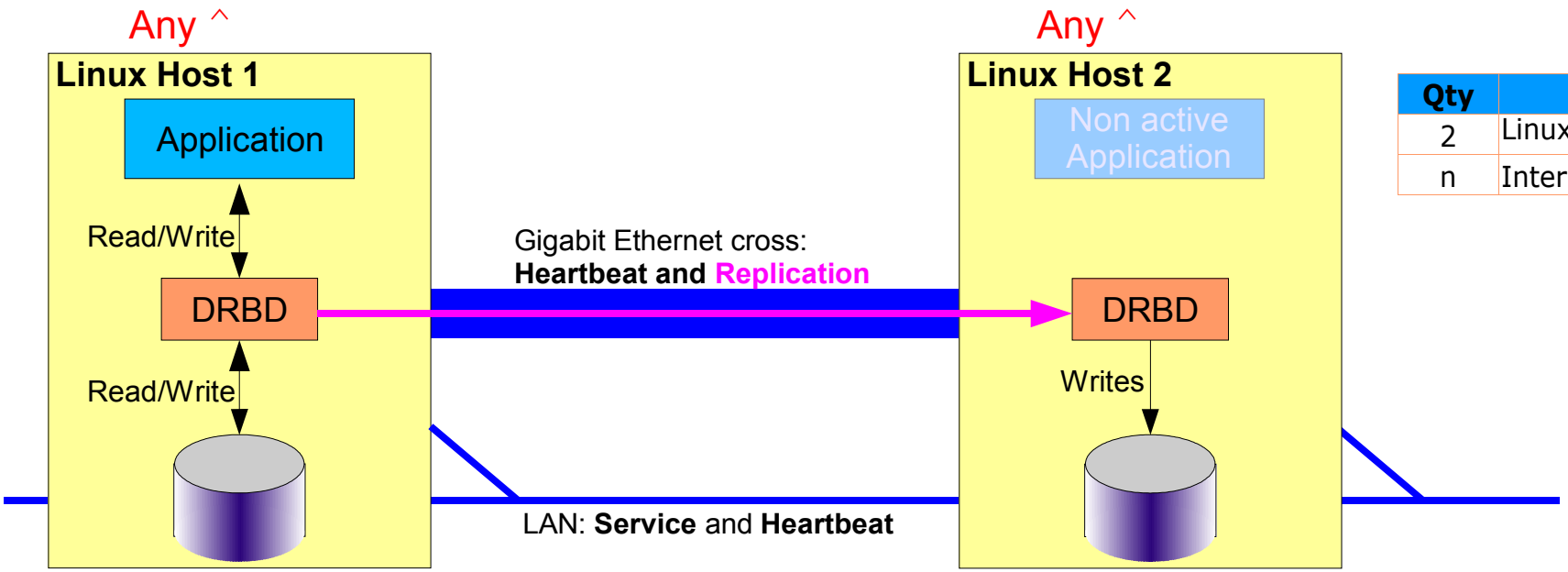
Qty	Description
2	~ xSeries
1	DS300 or DS400 disk array
n	SCSI disks
2	ServeRAID4* or 6M board
2	SCSI cable
1	Ethernet cross cable

- **Host 1** runs **AppA**. Optionally, **Host 2** may run **AppB**
- **Host 2** monitors availability of **Host 1** through some Heartbeat software
- In case of Active-Active cluster, **Host 1** may monitor **Host 2** simultaneously
- If **Host 1** fails, a set of automatic actions on **Host 2** will:
  - ✔ Take control of **Host-1**-owned disk arrays;
  - ✔ Configure **Host 2**'s network interfaces to respond as **Host 1** and **Host 2** simultaneously
  - ✔ Activate **AppA** on **Host 2**

ServeRAID and SCSI Clusters Details

- Support for Active-Passive and Active-Active clusters
- Red Hat Advanced Server's Cluster Manager is not supported with *clustered* ServeRAID
- Hardware feature: each logical array can be owned by one node at a time (this is why Red Hat's solution will not work)
- RAID 5 not supported for clustered environment, in any platform. See ServeRAID manual page 74
- Supported high availability software:
  - **System Automation for Multi Platforms (aka TSA)**
  - **Linux-HA**
  - **SteelEye's LifeKeeper**

Data Replication Clusters Overview



Qty	Description
2	Linux instances (or ^)
n	Internal disks

- **Host 1** runs **AppA**. Optionally, **Host 2** may run **AppB**
- **AppA** reads and writes to local storage via DRBD. Each byte written in **Host 1**'s local storage is replicated online to DRBD on **Host 2**.
- **Host 2** monitors availability of **Host 1** through some Heartbeat software
- In case of Active-Active cluster, **Host 1** can monitor **Host 2** simultaneously
- If **Host 1** fails, a set of automatic actions on **Host 2** will:
  - ✔ Activate **Host 2**'s replica of the storage (mount the volume);
  - ✔ Configure **Host 2**'s network interfaces to respond as **Host 1** and **Host 2** simultaneously
  - ✔ Activate **AppA** on **Host 2**

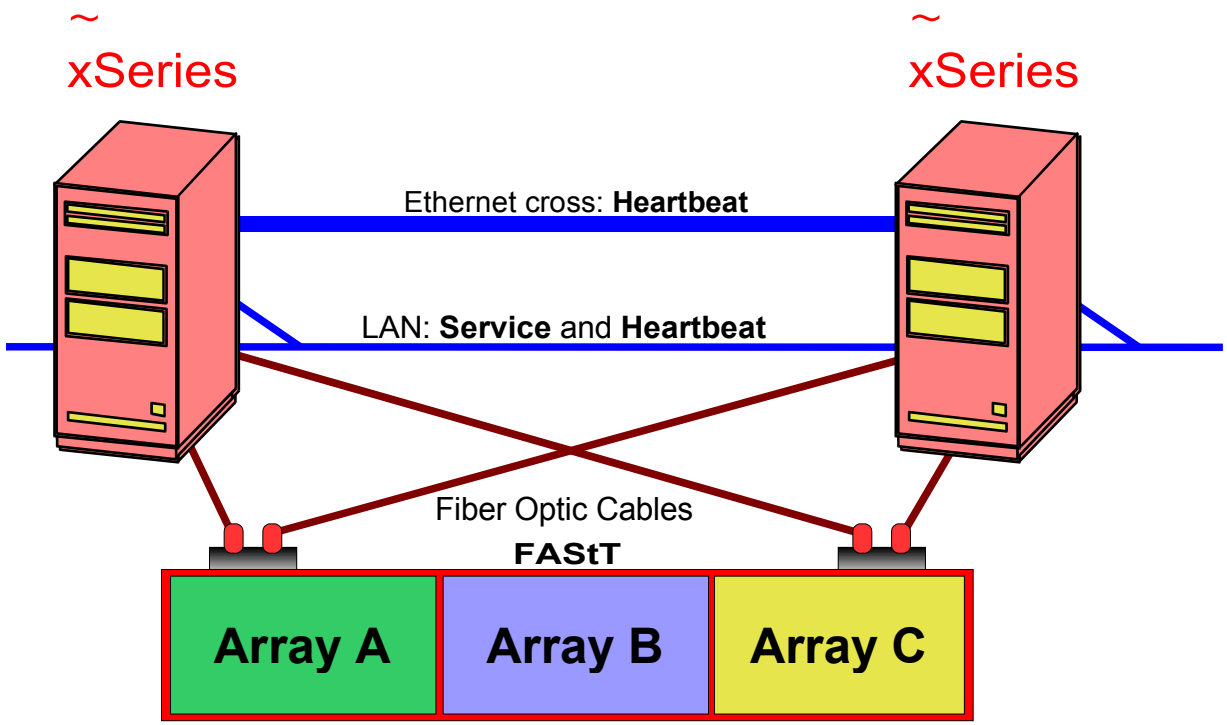
Data Replication Clusters Details

- DRBD stands for *Data Replication Block Device*. It is a low level replication software that operates on the OS level. *http://www.drbd.org/*
- DRBD is completely transparent to the application and users. No special support for replication is required in application
- Support for Active-Passive and Active-Active clusters
- Supports LVM and HW RAID. Linux on any ^ family, virtual Linux instances or mixed
- DRBD can be used for *cross-site replication*. SW RAID is a waste of time and CPU when DRBD is used
- Customer reference regarding replication bandwith: Hospital running Informix on a DRBD cluster with observed peak load of 6MBits/s over the 100MBits/s fast ethernet interface used for replication
- zSeries High Availability Red Paper: http://www.redbooks.ibm.com/abstracts/redp0220.html
- Supported high availability software:
  - **Linux-HA** for all platforms
  - **SteelEye's LifeKeeper** for ^ xSeries
  - DRBD is not compiled by default on **Red Hat AS**. But IGS can easily provide this functionality on RHEL

SAN Based Linux Clusters

- Fully shared storage clusters brings a new level of clustering to Linux
- Leverages high performance parallel applications
- While ServeRAID requires that each node keeps its own private logical array (transferable in a fail-over situation), SAN don't. Each storage byte is truly shared across nodes.
- When shared storage is needed, a software layer is required to manage simultaneous multiple nodes access. This is provided by things like **GPFS**, **OpenGFS**, **GFS**, **Oracle File System** (included in RHAS), etc.
- Due to higher price, SAN storage should be chosen when simultaneous data access is a must. Other way, SCSI may provide a better price/benefit relation.

Simplest SAN Cluster



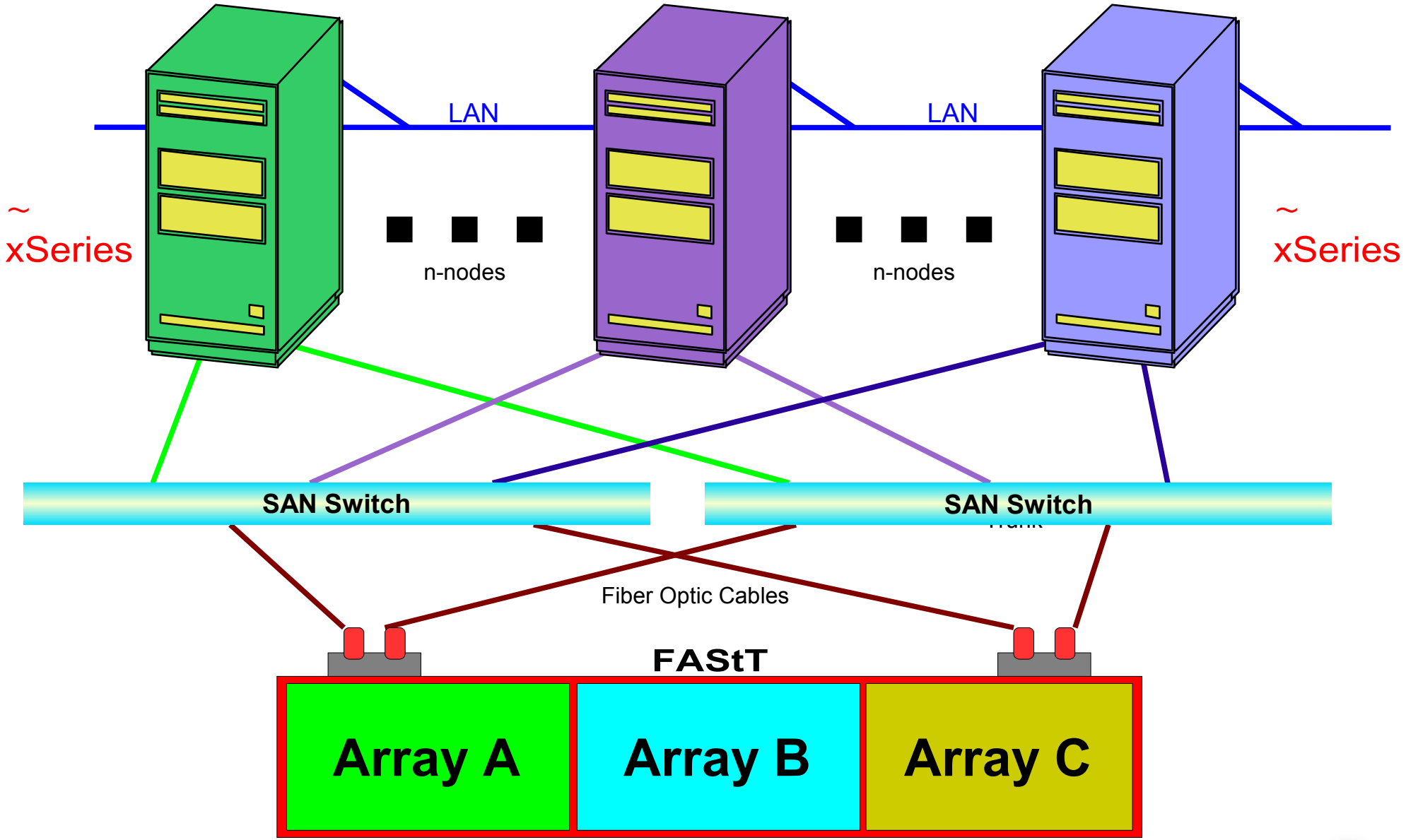
Qty	Description
2	~ xSeries
1	FAStT storage array
n	Disks
2 or 4	Fiber Channel board
2 or 4	Fiber cables
1	Ethernet cross cable

- One FC cable per node is also supported
- Both nodes may be active, running the same parallel application. But generally one node runs a special application process to balance requests: This is the Controller Process
- Some High Availability software is needed to move the Controlling Process from one node to another, when in a failure situation

SAN Clusters Details

- Support for Active-Passive and Active-Active clusters, but Parallel Active-Active applications takes more benefit from the HW.
- When each node has 2 FC boards, special software, provided with the product, is needed to run in each node
- FAStT support:
  http://www.storage.ibm.com/disk/fastt/supserver.htm
- Supported high availability software:
  - **System Automation for Multi Platforms (aka TSA)**
  - **Linux-HA**
  - **Red Hat Advanced Server**
  - **SteelEye's LifeKeeper**
  - **Polyserve**

Advanced SAN Cluster



Which Storage to Use

The best (price/perf) storage solution is 100% dependent on the application that will be hosted

Context / Storage Type	DRBD	ServeRAID/SCSI	FAStT/FC
Majority of HA clustered server applications	✔	✔	✔
Parallel application that requires multinode simultaneous shared storage	✘	✘	✔
Parallel application that does not require multinode simultaneous shared storage (like DB2)	✔	✔	✔
Storage consolidation with 3 or more nodes using same storage device	✘	✘	✔
HA Solution that needs simultaneous shared storage access (Like RH Cluster System)	✘	✘	✔
Customers that don't like Data Replication, can't deal with resync or have extremely (seriously extreme) write-intensive application		✔	✔

A Note About Storage Sizing

Oversized architectures are very common. Customers of new applications don't have a clue of their storage needs. Here are some real world examples.

1. Minimum Red Hat Linux installation (best for a server): **800MB**
2. Storage needed to install OS + DB2 + WebSphere or Oracle software files: **1GB**
3. Database size of one of brazilian biggest B2B portal running Ariba: **20GB**
4. General Motors in Brazil all portals (B2B, B2C) web content size: **2.8GB**
5. Same portals total database size on disk: **670MB**
6. Same portals compressed web server logs for 1 year: **40GB**
7. Bookseller B2B web content: **3GB**
8. DB size on disk of a brazilian online yellow pages service: **18GB**
9. Web content of same: **300MB**
10. Portal web content (EARs, JARs) of a small bank: **10MB**
11. Brazil's large SAP customer's HR database size: **60GB**
12. A big call center CRM database size: **5GB**
13. Oracle ERP DB size for a company in the automotive sector: **111GB**
14. *Single* SCSI HD as of 02H03: **140GB**

Source: IBM Strategic Outsourcing / Web Hosting delivery team and Customers. 02H03

Popular Software Details

DB2 on Linux Clusters

DB2 Information Management Software

- DB2 UDB parallel server on n-node Intel Linux Cluster is the more cost effective, high performance database. Intel hardware is a commodity, and horizontal scalability is virtually infinite
- Even when used as a parallel application, DB2 is a Share-Nothing database. No simultaneous shared storage like SAN is required. DS300/400 (SCSI) or even internal storage with replication may be used for inexpensive configurations
- Storage sizing for DB2 (or any database) must contemplate **backup** and **transaction log** sizes, not only the pure data size. Sizing must be done by the customer DBA that knows the application
- DB2 Parallel needs HA software to take-over the Controlling Process. SA/MP (aka TSA) is included for free on DB2 Parallel on Linux and AIX, for any number of nodes
- Detailed DB2 on DRBD Linux cluster configuration manual: http://www-3.ibm.com/software/data/pubs/papers/#db2halinux

Oracle RAC on Linux Clusters

- Oracle 9i Real Application Cluster is a complete solution for n-node Linux Clusters. It is quite popular today
- Oracle is a share-all database, so only FC configurations are supported
- Oracle filesystem (included in Red Hat Enterprise Advanced Server 3.0) is the software layer that will guarantee shared data integrity
- No HA software is needed. Oracle provides HA for the controlling process. Oracle on Red Hat AS + xSeries servers + SAN storage is a complete solution
- Storage sizing for DB2 (or any database) must contemplate **backup** and **transaction log** sizes, not only the pure data size. Sizing must be done by the customer DBA that knows the application
- Special cluster software may be used, but not necessary:
  - **Polyserve**

Lotus on Linux Clusters

Lotus. software

- From an HA Linux clustering perspective, Lotus Domino is a regular application with no parallel characteristics
- Domino provides database replication capabilities though, that can be used to eliminate shared storage and reduce solution price
- If external storage still needed, DS300/DS400 (SCSI) will do the job. No SAN is required

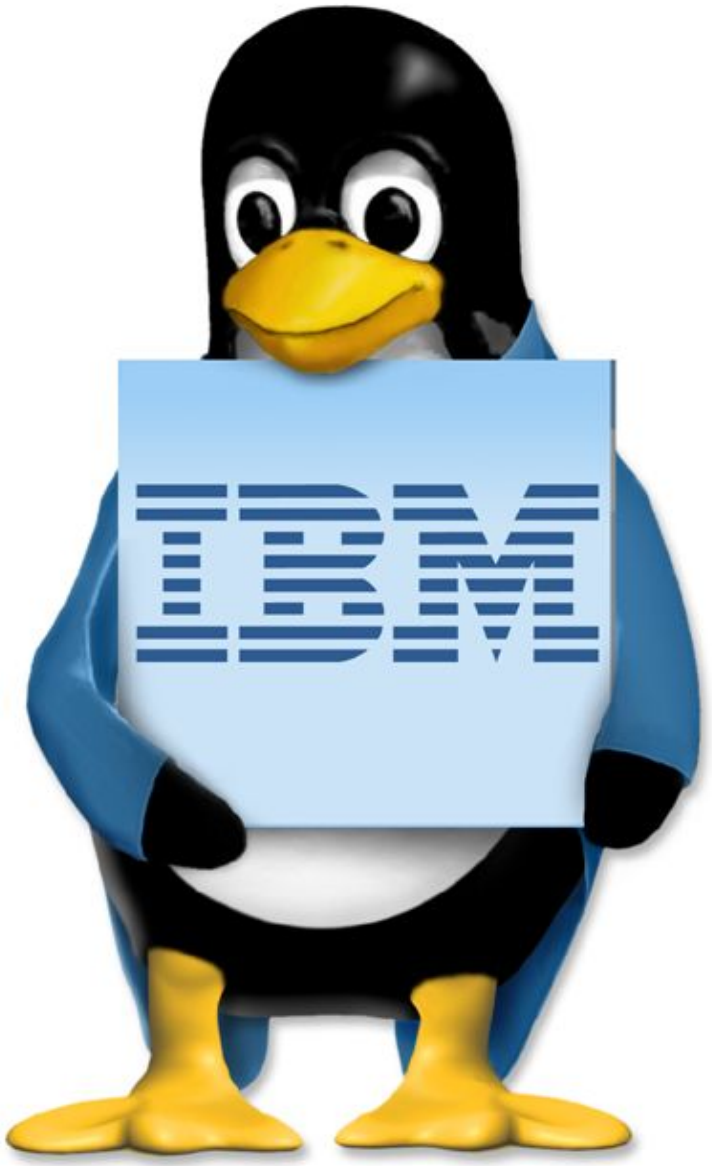
System Automation for Multiplatforms

- Also known as Tivoli System Automation, is the IBM strategic solution for High Availability
- Since version 2.1, supports Linux DRBD (replication), ServeRAID and Fiber Channel storages, due to its storage-agnostic features
- Powerful high level policy and rules creation with grouping and relationships
- Reduce implementation time, coding and support effort with Automation Policy
- Can be integrated in a Tivoli-managed data center
- HA infrastructure derived from RSCT, from IBM AIX's HACMP
- Mainframe-like High Availability for Linux, toward autonomic end-to-end automation
- Currently supported on ^ xSeries, pSeries, AIX and zSeries

Tivoli software

HA Linux Clusters Contacts

- Mark Price/Beaverton/IBM@IBMUS
  IGS L3 contact (only if local IGS contracted)
- Alan Robertson/Denver/IBM@IBMUS
  LTC HA Linux Clusters specialist
- Carol Carson/Poughkeepsie/IBM@IBMUS
  xSeries Linux Sales Enablement
- Douglas McGuire/Lexington/IBM@IBMUS
  Americas Linux Clusters
- Joachim Schmalzried/Germany/IBM@IBMDE
  IBM System Automation for Multiplatforms
- High Availability Software:
  - **Tivoli System Automation:** http://ibm.com/software/tivoli/products/sys-auto-linux
  - **Linux-HA**: http://www.linux-ha.org
  - **Red Hat Advanced Server:** http://www.redhat.com
  - **SteelEye's LifeKeeper**: http://www.steeleye.com
  - **Polyserve:** http://www.polyserve.com



Avi Alkalay <avix@br.ibm.com>

+55-11-2132-2327

Linux, Open Standards Consultant
IBM Corporation

Thank You !

People on This Document

- Alan Robertson: HA Linux Clusters LTC specialist
- Alcino Bras: FAStT and SAN info
- João Marcos: DB2 review
- Moacir Malemont: Lotus review
- Joachim Schmalzried: Tivoli System Automation awareness
- Jeferson Moia: xSeries part numbers
- Ana Maria Bezerra: Strategic Outsourcing real-world sizing numbers