NFS INDUSTRY CONFERENCE

Global Storage Architecture

Scalable NFS service using off the shelf components

Stanley Wood

Senior Software Engineer

IBM

swood@us.ibm.com

Special Notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government customers. Rates are based on a customer's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Many of the pSeries features described in this document are operating system dependent and may not be available on Linux. For more information, please check: http://www.ibm.com/servers/eserver/pseries/linux/whitepapers/linux_pseries.html.

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.

Special Notices (Cont.)

The following terms are registered trademarks of International Business Machines Corporation in the United States and/or other countries: AIX, alphaWorks, AS/400, Blue Lightning, C Set++, CICS, CICS/6000, ClusterProven, DataHub, DataJoiner, DB2, DEEP BLUE, DFDSM, DirectTalk, DYNIX, DYNIX/ptx, ebusiness(logo), ESCON, FlashCopy, GDDM, IBM, IBM(logo), IntelliStation, IQ-Link, LANStreamer, LoadLeveler, Magstar, MediaStreamer, Micro Channel, MQSeries, Net.Data, Netfinity, NetView, Network Station, NUMA-Q, Operating System/2, Operating System/400, OS/2, OS/390, OS/400, Parallel Sysplex, PartnerLink, PartnerWorld, POWERparallel, PowerPC, PowerPC(logo), ptx, ptx/ADMIN, RISC System/6000, RS/6000, S/390, Scalable POWERparallel Systems, SecureWay, Sequent, ServerProven, SP1, SP2, System/390, The Engines of e-business, ThinkPad, Tivoli, Tivoli(logo), Tivoli Management Environment, Tivoli Ready(logo), TME, TURBOWAYS, VisualAge, WebSphere.

The following terms are trademarks of International Business Machines Corporation in the United States and/or other countries: ~,AIX/L, AIX/L(logo), AIX 5L, AIX PVMe, AS/400e, Blue Gene, Chipkill, C2T Interconnect, DB2 OLAP Server, DB2 Universal Database, DFSORT, e(logo)business, e-business(logo), e-business on demand, eLiza, Enterprise Storage Server, GigaProcessor, HACMP/6000, IBMlink, IMS, Intelligent Miner, iSeries, Light Path Diagnostics, NUMACenter, PowerPC Architecture, PowerPC 604, POWER2, POWER2 Architecture, POWER3, POWER4, POWER4+, pSeries, Sequent (logo), SequentLINK, Server Advantage, Service Director, SmoothStart, SP, Tivoli Enterprise, TME 10, TotalStorage, Ultramedia, Videocharger, Visualization Data Explorer, X-Architecture, xSeries, zSeries.

A full list of U.S. trademarks owned by IBM may be found at: http://www.ibm.com/legal/copytrade.shtml.

Lotus, Lotus Notes, Lotusphere and Notes are registered trademarks and Domino is a trademark of IBM-Lotus in the United States and/or other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group.

Microsoft, Windows, Windows NT and the Windows logo are registered trademarks of Microsoft Corporation in the United States and/or other countries.

Intel and Itanium are a registered trademarks and MMX, Pentium and Xeon are trademarks of Intel Corporation in the United States and/or other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States and/or other countries.

Other company, product and service names may be trademarks or service marks of others.

Global Storage Architecture

Background

GSA Objectives

- Replace legacy file systems infrastructure
  - AFS/DFS
  - OS/2 and Windows
  - Standalone NFS/NAS
- Support varied user population
  - Hardware design
  - Software development
  - Office productivity

GSA Requirements

- Highly scalable

  Some legacy sites support over 1 billion transactions per day

- Highly reliable

  Even brief outages impact design work

- Less expensive than legacy environments
- Consistent service and policies

  Same user experience world wide

GSA Requirements

- Future proof architecture
  - No more migrations
- Same or better performance
  - Custom benchmarks designed by GSA users
- Use industry standards
  - Wherever possible

GSA Requirements

- Customers and administrators expect features from AFS/DFS
  - Global namespace
  - Location independence
- We wanted to avoid
  - Thousands of file system mounts
  - Excessive use of symbolic links
  - Fragmented userid space

Global Storage Architecture

Architecture

GSA Architecture

- NFS and SMB/CIFS obvious choices for file system protocols
  - Industry Standard/Ubiquitous
  - No proprietary client software required
- How do we deliver on requirements and expectations with NFS and CIFS?

GSA Architecture

- Clustered design gives many benefits of legacy systems while using commodity components
  - Scalability
  - Reliability
  - Location independence
  - Global namespace
  - Replication

GSA Architecture

- A scalable, robust fileserver using off the shelf components
  - General Parallel File System (GPFS)
    - Clustered file system
  - AIX pSeries servers
    - NFS file service
  - WebSphere Edge Server/Network Dispatcher
    - Load balancer

GSA Architecture

- off the shelf components (Cont.)
  - Apache
    - Web access to file system
    - Management tools
  - Samba
    - Windows access
  - OpenLDAP
    - Directory software
  - ProFTPD
    - FTP access to file system

General Parallel File System

- GPFS is the core file system in GSA
- Designed for multimedia services
  - Massive throughput
  - High reliability
- Cluster file system
  - Shared device file system
  - Exploits SAN attached storage devices
- Provides read/write data replication
- Runs on AIX and Linux

Service Delivery Agents

- Most of the servers in GSA are SDAs
- pSeries servers running AIX
- Client systems connect to SDAs
  - Connect with client system's native protocol
  - The SDA connects to GPFS
- All SDAs are exactly alike
  - GSA cells can survive SDA failures
  - GSA cells can scale up or down seamlessly by adding or removing SDAs and SAN storage

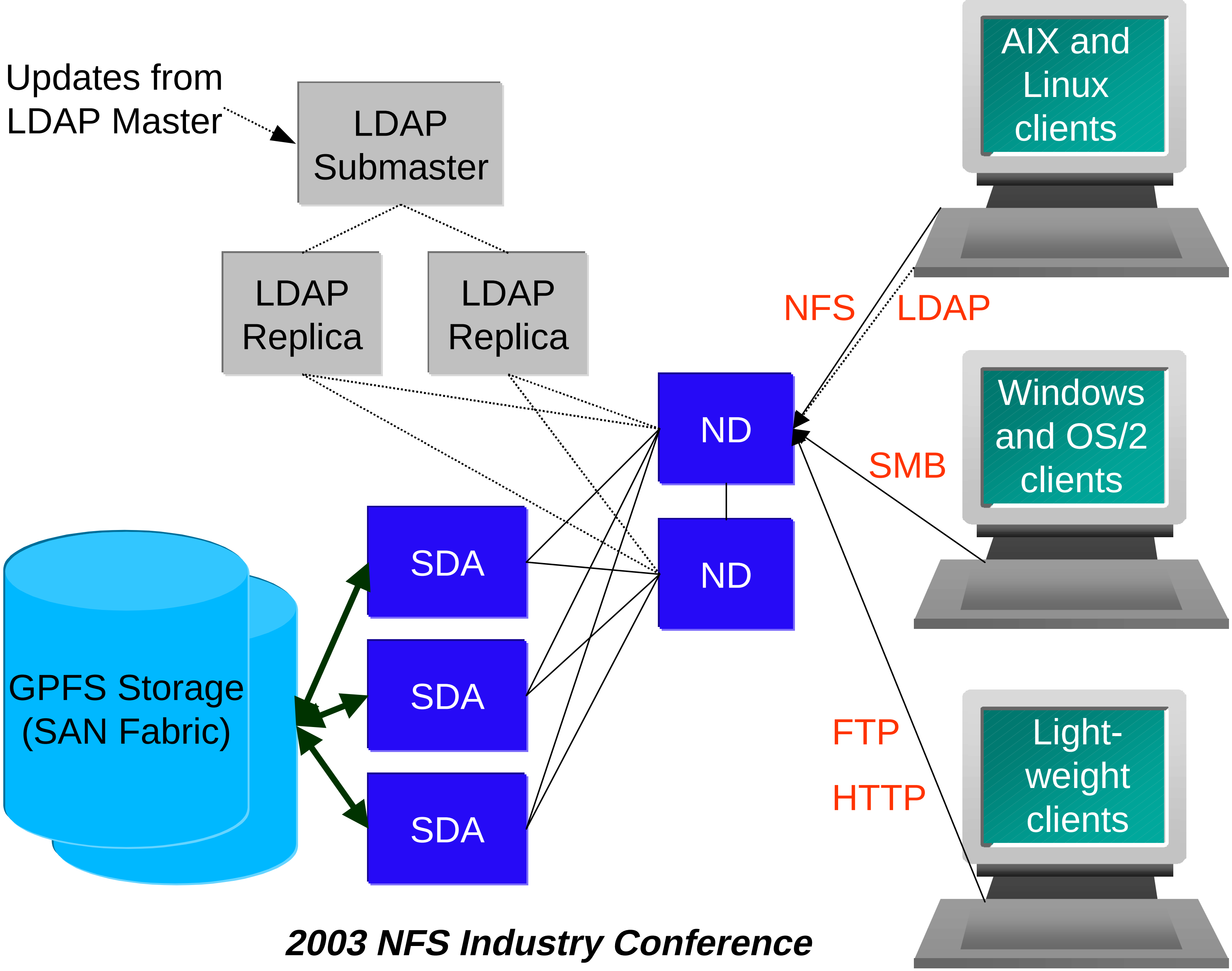
Directory Servers

- Keep common information across cells
  - Users and user information
  - Groups and group membership
  - Automount maps and cell information
- xSeries servers running Linux
- Directory is replicated world wide
  - Single master server accepts all updates
  - Updates are pushed through a replication hierarchy to replicas in every cell

Network Dispatchers

- Load balancing component of WebSphere Edge Server
- Manages all network access to GSA
  - Inbound packets go through network dispatchers
  - Outbound packets bypass dispatchers
- New connections routed to best server
  - LDAP or SDA back end as appropriate
  - Choose the most responsive server
- Mask server failures and maintenance

GSA Cell Architecture



Updates from LDAP Master

LDAP Submaster

LDAP Replica

LDAP Replica

ND

ND

SDA

SDA

SDA

GPFS Storage (SAN Fabric)

NFS

LDAP

SMB

FTP

HTTP

AIX and Linux clients

Windows and OS/2 clients

Light-weight clients

GSA Global Namespace

- Cells all share common LDAP directory
  - Users have the same ID in every cell
  - All groups exist in every cell
- No need for intercell constructs
- Cells are addressable consistently
  - Web: http://<region>gsa.ibm.com/<path>
  - FTP: ftp://<region>gsa.ibm.com/<path>
  - Windows: \\<region>gsa.ibm.com\<path>
  - NFS: /gsa/<region>gsa/<path>

GSA File System Layout

- Each GSA cell contains four top level directories where users can request space:
  - /home
  - /projects
  - /system (shared tools)
  - /tdisk (temporary space)
- Hash trees minimize the number of subdirectories per directory
  - /home/<letter>/<letter>/
  - /projects/<letter>/
  - /tdisk/<date stamp>/

GSA File system Exports

/gsa/pokgsa	Hash trees and system
/gsa/pokgsa-h1	Home directories
/gsa/pokgsa-h2	Home directories
/gsa/pokgsa-p1	Project directories
/gsa/pokgsa-p2	Project directories
/gsa/pokgsa-t1	Temporary directories

GSA Space Management

- Physical Space
  - GPFS supports up to 32 file systems per cluster
  - Each file system can grow to many terabytes
  - File systems can grow and shrink dynamically
  - Automated processes adjust file system sizes
- Quotas
  - GPFS supports quotas per user or group
  - GSA uses groups and group quotas to assign quotas to each project and home directory
  - Users are charged monthly for their average daily space usage

NFS INDUSTRY CONFERENCE

Global Storage Architecture

Status

GSA Performance

- GSA is performing on par with legacy services
  - Home grown performance tests
  - Better than AFS and DFS on 3 of 4 cases
- Additional work underway to improve performance
  - Overall tuning
  - Enhancements to GSA components
  - Expect to beat legacy systems in future

GSA Resources

- GSA is proving less costly to operate than the legacy services
  - Space rates down about 30%
  - Fewer admins required per gigabyte
  - Less floor space/gigabyte
- Results from
  - Greater automation
  - Ability to manage space in larger pools
  - Operating servers closer to capacity

GSA Status

- GSA entered pilot late in 2001
- Saw limited production in 2002
- Became a rated service in 2003
- GSA space and growth
  - 13 TB of used space in September 2003
  - Just over 1% of total IBM file system space
  - Monthly growth rate around 20%
  - 23,000 active user accounts

GSA Status

- 12 GSA cells in production

– Austin, TX	ausgsa.ibm.com
– Beaverton, OR	bvrgsa.ibm.com
– Burlington, VT	btvgsa.ibm.com
– Ehningen, Germany	ehngsa.ibm.com
– La Gaude, France	lgegsa.ibm.com
– North Harbor, UK	nhbgsa.ibm.com
– Poughkeepsie, NY	pokgsa.ibm.com
– Raleigh, NC	rtpgsa.ibm.com
– Rochester, MN	rchgsa.ibm.com
– San Jose, CA	snjgsa.ibm.com
– Yamato, Japan	jpngsa.ibm.com
– Yorktown Heights, NY	watgsa.ibm.com

GSA Status

- New cell deployments underway
  - Armonk
  - Dublin
  - Rome
- . . . with more to follow . . .
  - Boulder
  - Haifa
  - Toronto
  - et. al.

NFS INDUSTRY CONFERENCE

Global Storage Architecture

Questions?