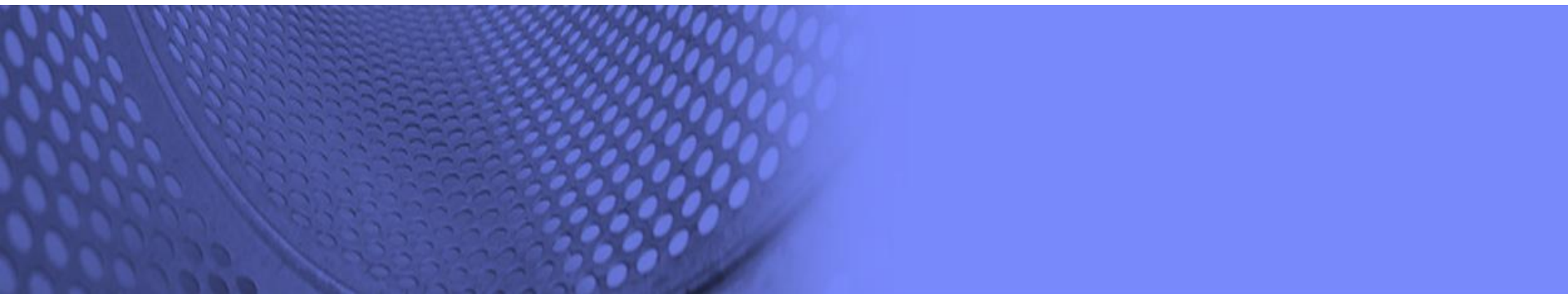# High performance and scalable central data repository for academic environments

Rudolf Hruska
Information Infrastructure Leader
IBM Systems & Technology Group
rudolf_hruska@cz.ibm.com

IBM

# Agenda

**Why we built Scale Out NAS (SONAS)**
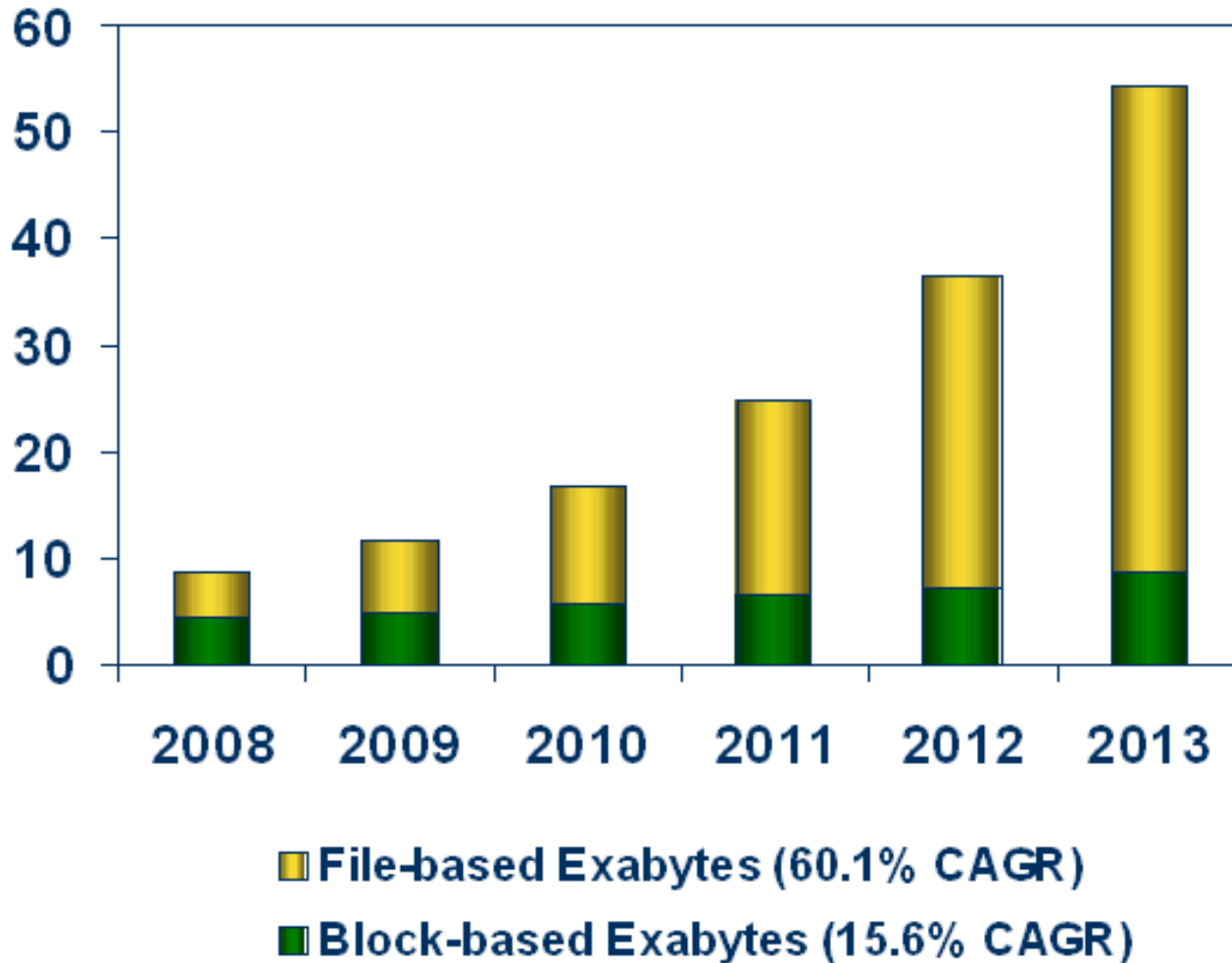
_____

**Policy-based Data Management**

_____

**The SONAS Architecture**

_____

**Outlook**

# Globally, storage requirement  is *80% file-based unstructured data*, and growing

**Worldwide Storage Capacity Shipped by Segment, 2008–2013**



■ **File-based Exabytes (60.1% CAGR)**

■ **Block-based Exabytes (15.6% CAGR)**

*Source: IDC, State of File-Based Storage Use in Organizations: Results from IDC's 2009 Trends in File-Based Storage Survey: Dec 2009: Doc # 221138*

# Challenge 1 :  Proliferating NAS & File Servers

- Growing # of systems, of administrative effort, power consumption, etc.

**Global Namespace**

**+3 years**

**+3 years**

**File-level Virtualization**

leads to...
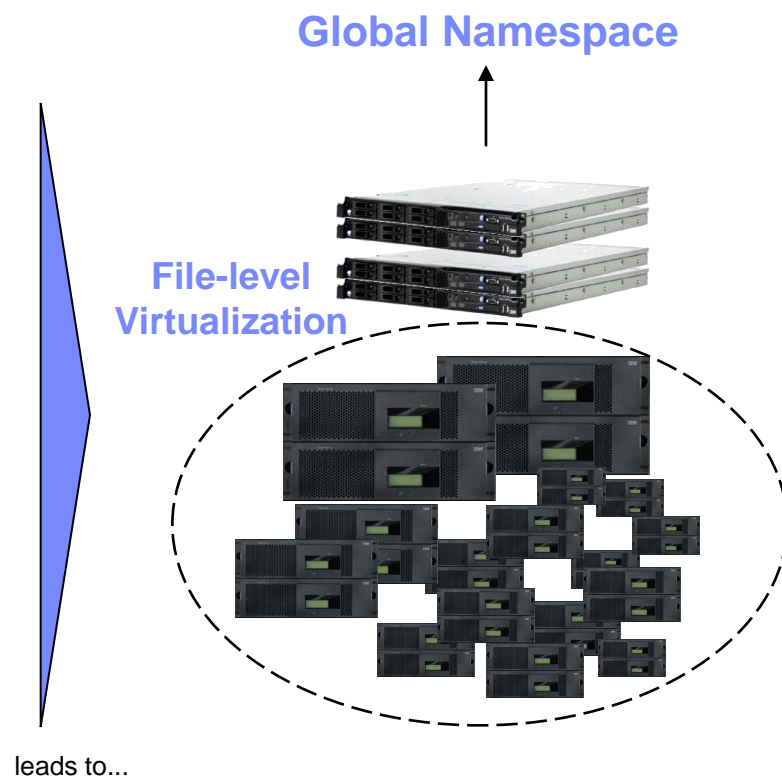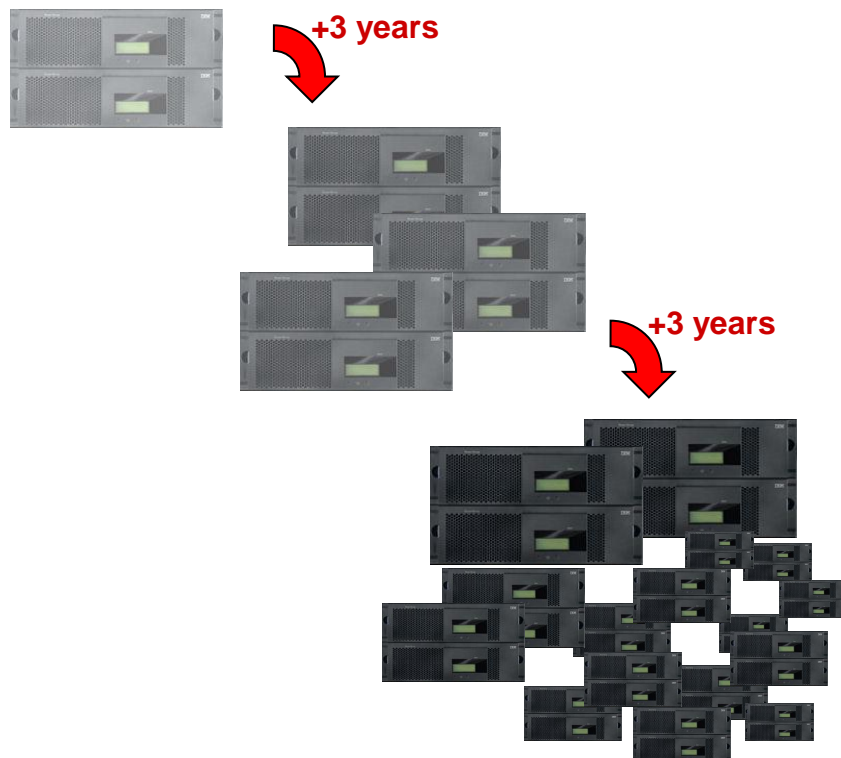
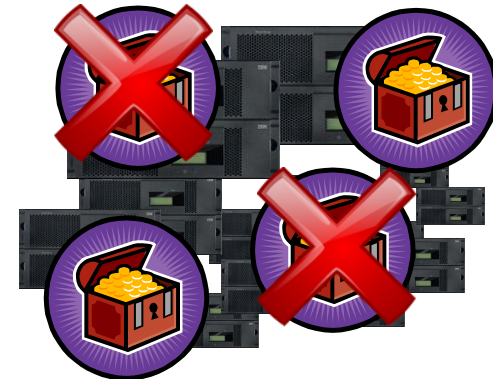# Challenge 2 : Users don't clean up

- **Userspace** consists of 50% or more inactive data. Just which 50%?
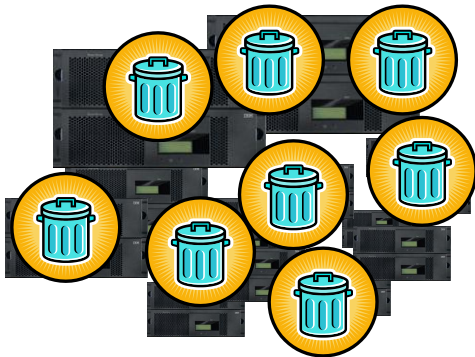
- **Tiered Storage**: Several GOLD projects have become obsolete; should turn back tier1 resources

# Challenge 3 : Manual Tiering gets cumbersome

- …True for data lifecycle management  *(how important?)*

- ... Also true for performance management  *(maximize throughput)*

# Introducing SONAS

# SONAS: Manage Data independently from involved Hardware

- **driven by content metadata policies** (name, type, date, age, frequency of use…)

- **infinitely scalable**

Global Name Space

**SONAS**

Policies
Policies

**Quota**

**Access Rights**

**Lifecycle / Copies / Disposal**

# Use "Overflow" Policy to achieve 90% Fill Grade

- Maintain high fill grade on gold tier while minimizing risk

- Filesystem Policy Automation – no daemons, no scripts

❶    **Fill grade > 90% → downtier unused files**

❷    **Continually premigrate stale files**

❸    **Files not used > 1 year → keep only *stub file***

**90%**

TSM Archive Overflow

# Scientific Background of GPFS Geo-Distribution Caching

## Panache: A Parallel File System Cache for Global File Access

Marc Eshel      Roger Haskin      Dean Hildebrand      Manoj Naik      Frank Schmuck
Renu Tewari
*IBM Almaden Research*
{eshel, roger, manoj, schmuck}@almaden.ibm.com, {dhildeb, tewarir}@us.ibm.com

### Abstract

Cloud computing promises large-scale and seamless access to vast quantities of data across the globe. Applications will demand the reliability, consistency, and performance of a traditional cluster file system regardless of the physical distance between data centers.

Panache is a scalable, high-performance, clustered file system cache for parallel data-intensive applications that require wide area file access. Panache is the first file system cache to exploit parallelism in every aspect of its design—parallel applications can access and update the cache from multiple nodes while data and metadata is pulled into and pushed out of the cache in parallel. Data is cached and updated using pNFS, which performs parallel I/O between clients and servers, eliminating the single-server bottleneck of vanilla client-server file access protocols. Furthermore, Panache shields applications from fluctuating WAN latencies and outages and is easy to deploy as it relies on open standards for high-performance file serving and does not require any proprietary hardware or software to be installed at the remote cluster.

In this paper, we present the overall design and imple-

Traditionally, NFS (for Unix) and CIFS (for Windows) have been the protocols of choice for remote file serving. Originally designed for local area access, both are rather "chatty" and therefore unsuited for wide-area access. NFSv4 has numerous optimizations for wide-area use, but its scalability continues to suffer from the "single server" design. NFSv4.1, which includes pNFS, improves I/O performance by enabling parallel data transfers between clients and servers. Unfortunately, while NFSv4 and pNFS can improve network and I/O performance, they cannot completely mask WAN latencies nor operate during intermittent network outages.

As "storage cloud" architectures evolve from a single high bandwidth data-center towards a larger multi-tiered storage delivery architecture, e.g., Nirvanix SDN [7], file data needs to be efficiently moved across locations and be accessible using standard file system APIs. Moreover, for data-intensive applications to function seamlessly in "compute clouds", the data needs to be cached closer to or at the site of the computation. Consider a typical multi-site compute cloud architecture that presents a virtualized environment to customer applications running at multiple sites within the cloud. Applications run inside a virtual machine (VM) and

# GPFS has Field Experience

- Supercomputing basis technology

**GPFS**

~50% of supercomputing capacity

- **Parallelization** = high scalability without Hotspots

- **Virtualization** = best yield

# What is in SONAS

- "GPFS-in-a-box" NAS with LDAP and TSM-HSM integration

- Extensive automation

- Maximum density

- 27 TB to 14,4 PB

- 2 to 30 interface nodes

- 60 to 7200 disk drives, SATA+SAS
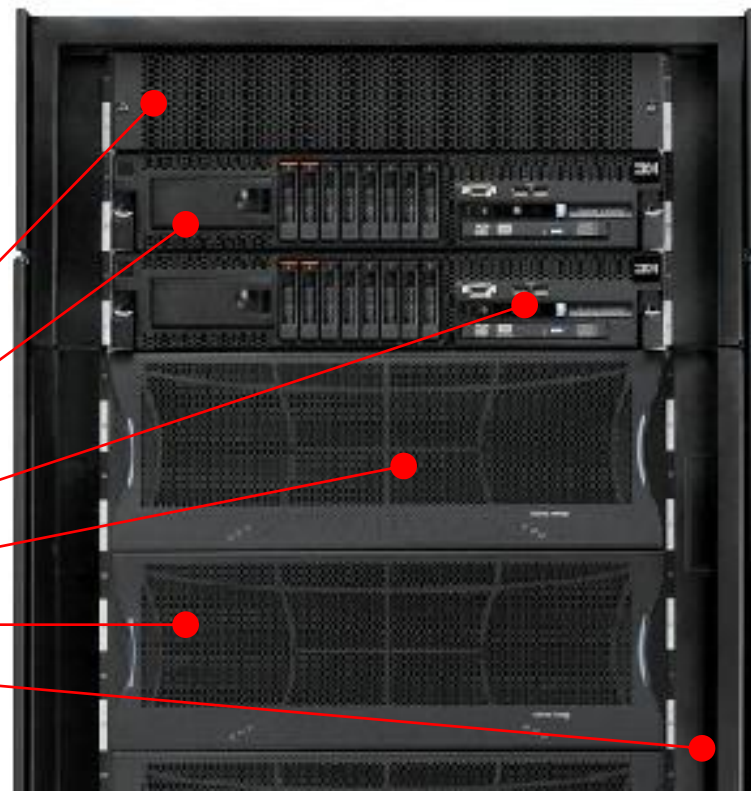
- 1 to 16 pre-built racks, rigid layout

- InfiniBand interconnect
  (2  10 Gbps per node)

**27, 60, 120 TB**
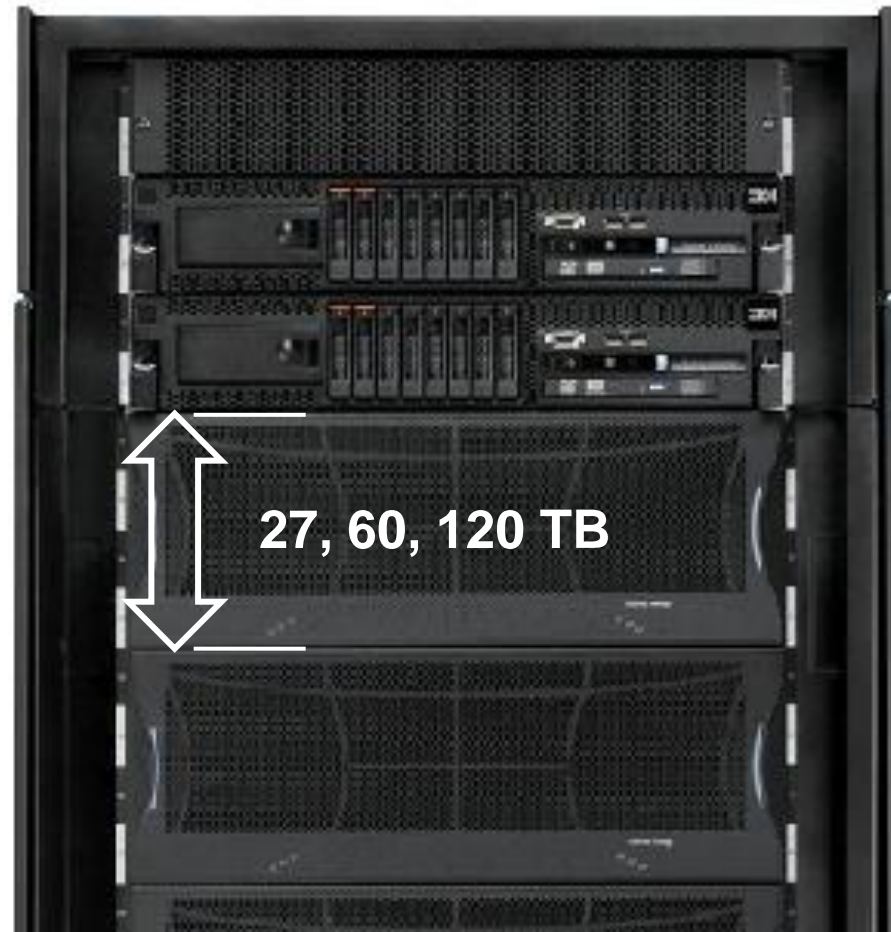
# SONAS versus IBM GPFS Clusters :  The Big Differences

- *Storage Product* with IBM product liability:  compliance, certifications, …

- **All** supported configurations are pre-tested

- Extensive added RAS functions + call home

- GPFS & InfiniBand are not exposed

- Standardized maintenance procedure

- One consolidated support process

Image: Voltaire

# SONAS is based on 2-Tier Scale-Out Architecture

User LAN

NFS CIFS FTP HTTP SCP

| I/O Node | I/O Node | I/O Node | I/O Node | I/O Node | ▪▪▪ max. 30 |

GPFS    **InfiniBand**    **Ethernet**    misc

**StoragePod**    **StoragePod**

▪▪▪ **max. 30,**  each 240 disks

**storage pool**  (SATA 1TB/2TB, FC 450GB)

IBM
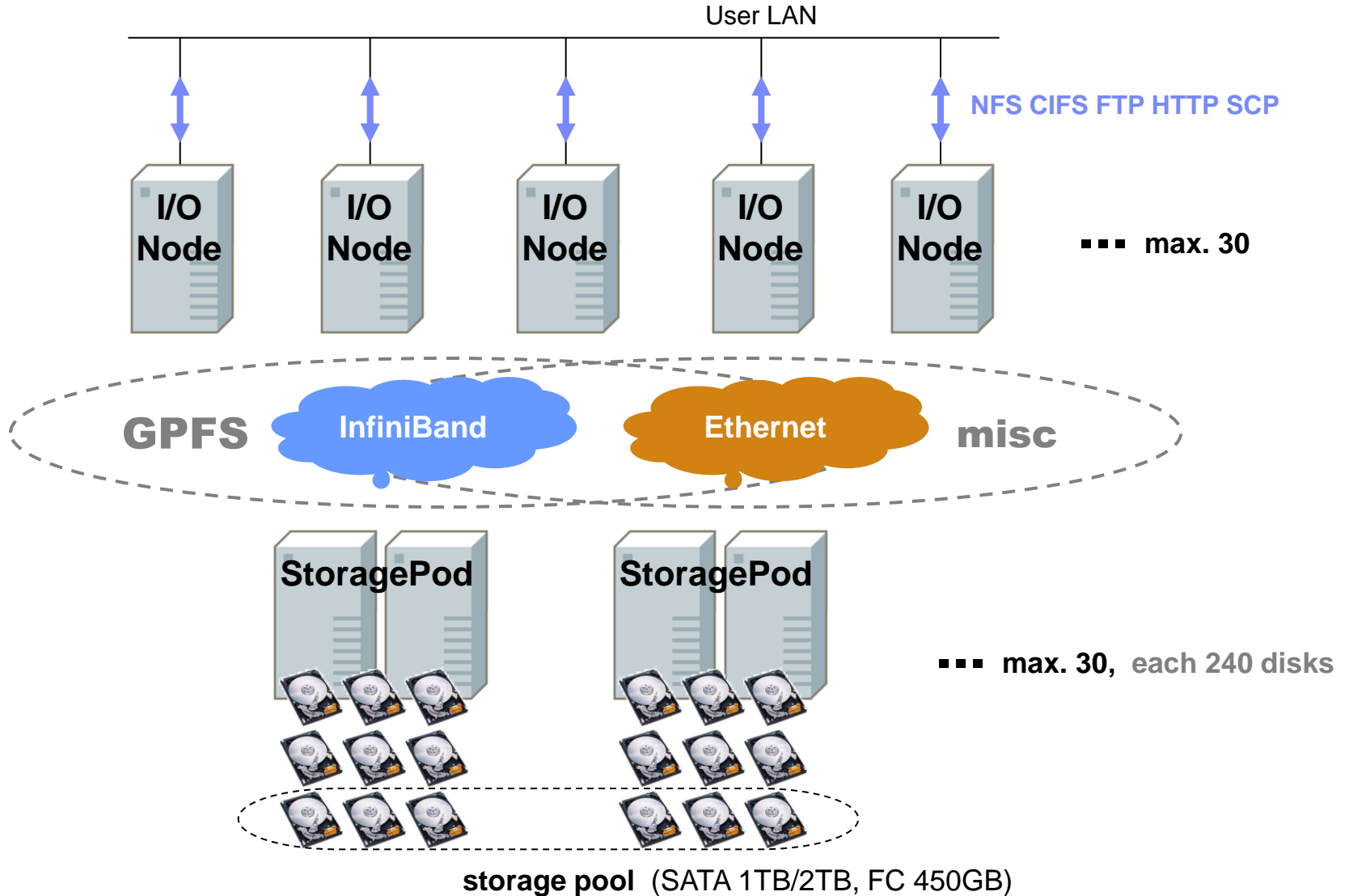
# Cache, Disks and Tape :  *One* Storage Continuum

Parallel  Windows® access
Uniform Windows/Unix view
Differential Snapshots
Wide Data Striping
**Quick Restore**

The whole storage stack including cache (RAM)
and backup/archive (tape) is integrally managed.

CIFS NFS

CIFS NFS

CIFS NFS

CIFS NFS

**Node**

**Node**

**Node**

**Node**

**Cache**

**Cache**

**Cache**

**Cache**

**Tape**  ◎ ◎  **TSM**
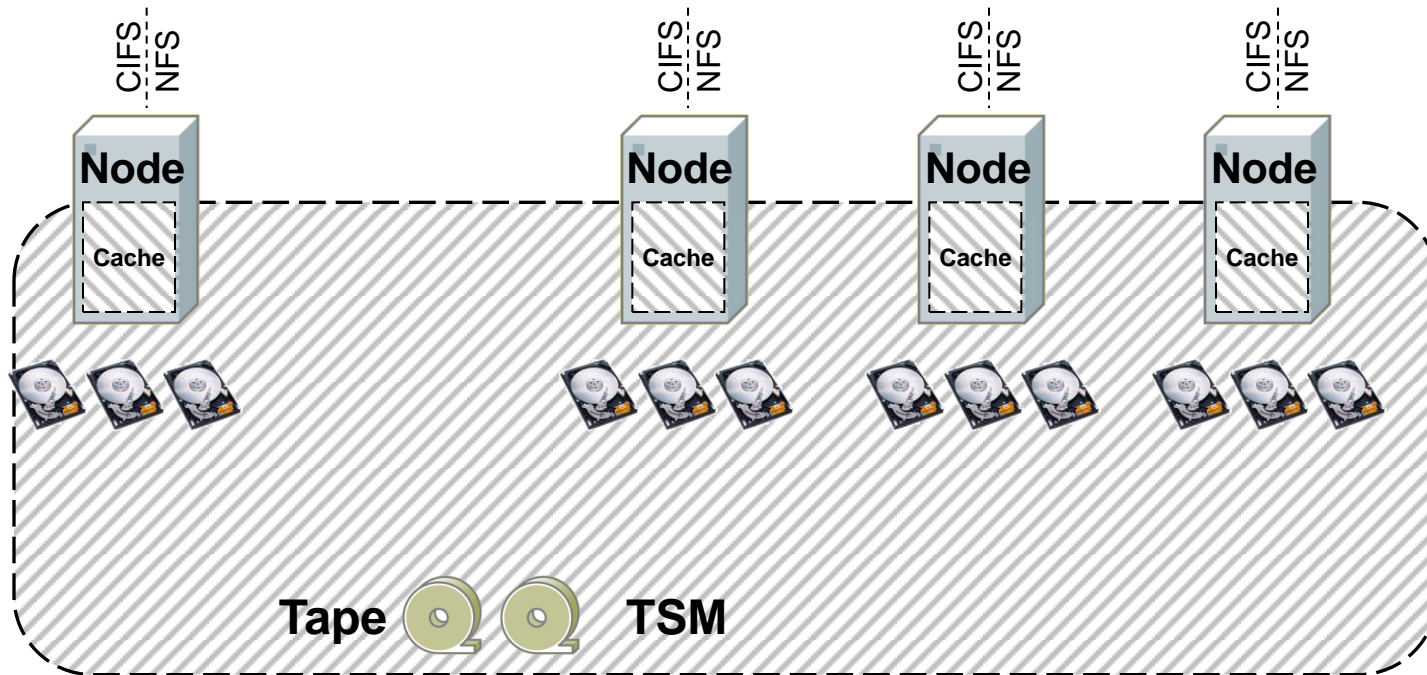
# SONAS Automated Lifecycle Management

More Examples – Policy-based, transparent for the user

CIFS NFS
**Node**

CIFS NFS
**Node**

CIFS NFS
**Node**

CIFS NFS
**Node**

Tape

**Archive stale data**

FC tier

**Data type *'db'* mirrored across two pools**

SATA tier

**Data type 'temp' always on tier 2**

**Overflow into *cloud***
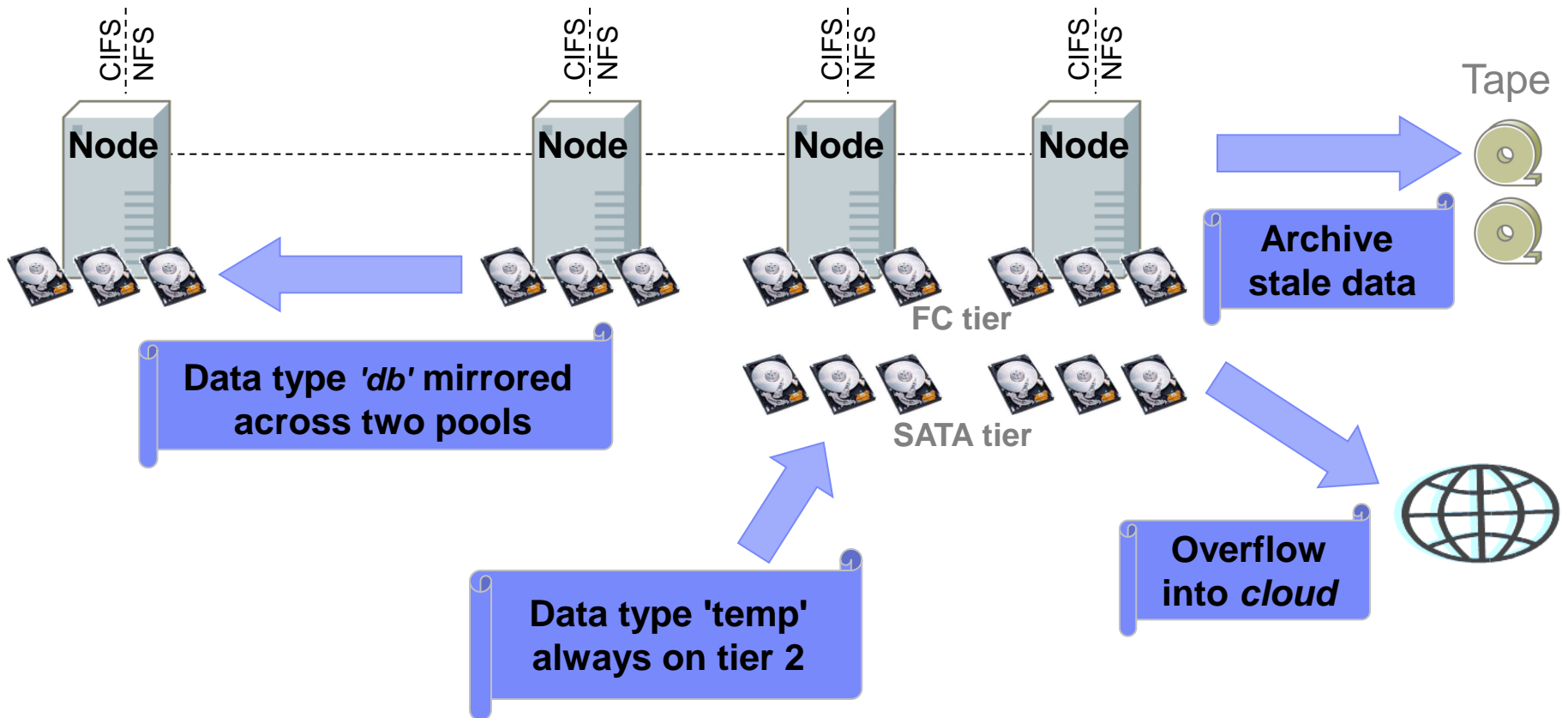
# Data Protection/High Availability Features

- **Snapshots**
    - Space efficient, differential snapshots
    - Includes Microsoft Windows VSS integration

- **Synchronous Replication**
    - File, Set of files or entire file system
    - Single site in initial release

- **Asynchronous Replication (Release 1.1.1)**
    - Any file system sub-tree from one cluster to another
    - Batched based, hub and spoke

- **High Availability (HA) Features**
    - Redundant Interface Nodes: allows access to data by users
    - Redundant Storage Nodes: allows access to storage
    - Redundant private 1GbE internal management network
    - Redundant private DDR Infiniband data network
    - RAID 5/6: Protects against individual disk failures
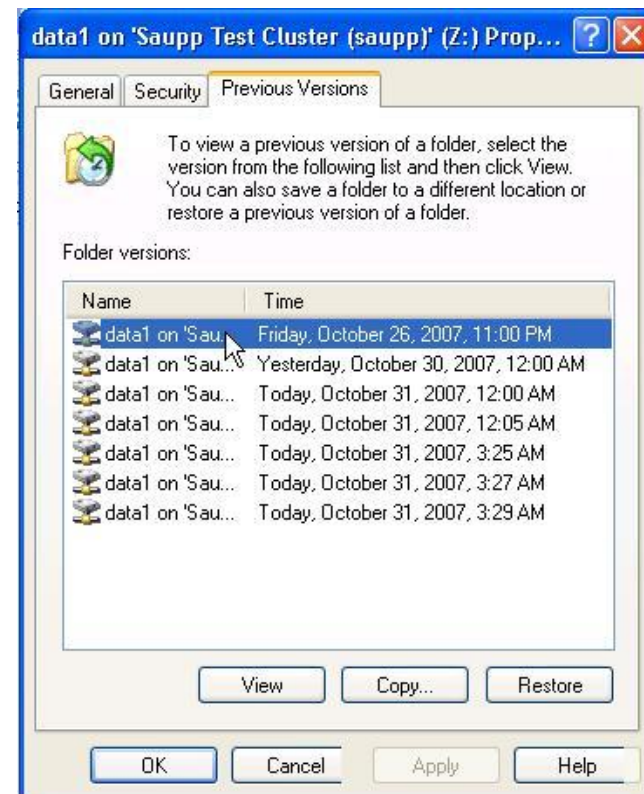
**Snapshots** Integrated into Windows Explorer using the Volume Shadow Copy Services (VSS)

**Snapshots** in Linux/Unix → `/.snapshots`

# SONAS System Utilization Monitor
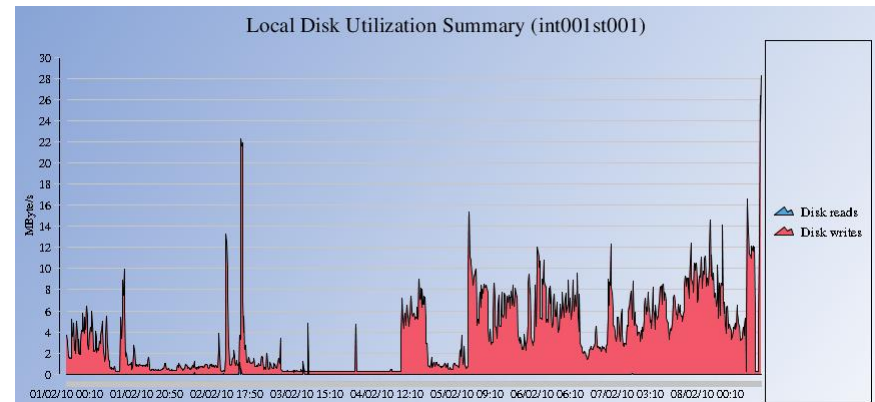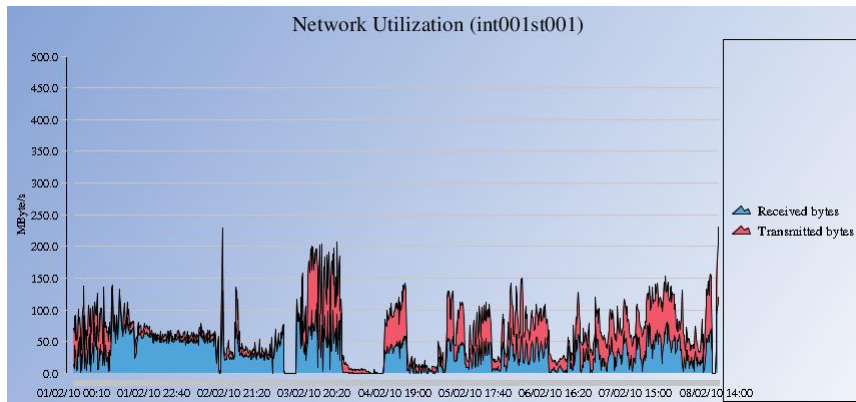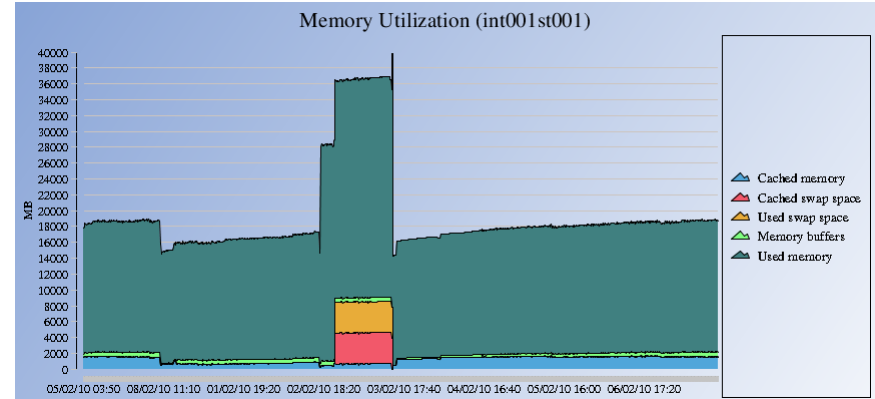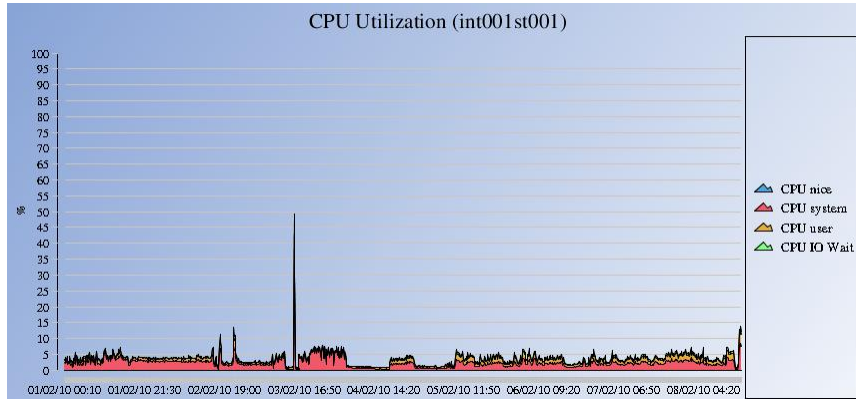
# SONAS Scalability & Summary

# SONAS Features at a glance

- **Network File Serving**
  - NFS v2/v3/v4*, CIFS, FTP, HTTP, SCP
  - CIFS ACL mapping into NFSv4 ACL
  - Coherent file locking between NFS and CIFS

- **Clustered parallel file system**
  - Up to 256 file systems
  - Up to 2 billion files per file system
  - Maximum 2PB per file system

- **Quota**
  - User, group and fileset level quotas
  - Soft limits, hard limits, grace periods

- **User Authentication/Authorization**
  - Microsoft® Active Directory
  - Lightweight Directory Access Protocol (LDAP) / with Kerberos
  - Samba primary domain controller (PDC)

- **Data Protection**
  - File system Snapshots, up to 256 per file system
  - Synchronous replication of file system metadata and file data
  - Integrated TSM V6.1 Backup/Archive (B/A) client

- **Centralized Management and Administration**
  - Both Graphical User Interface and Command Line Interface
  - Centralized alert log and event log
  - Event notifications via email or SNMP

- **Integrated Solution Packaging**
  - Single software product, multiple expandable hardware
  - All components integrated into rack(s), cabled, fully tested
  - Updates/patches via centralized SONAS patch management

- **Scalability and Performance**
  - ✓ Up to 30 interface nodes for I/O performance (Release 1)
  - ✓ High Density packaging of HDDs
  - ✓ Support for high performance 15K SAS disk drives and high capacity 7.2K SATA disk drives
  - ✓ Up to 7200 HDDs in single system (14.4PB using 2TB SATA)

- **RAS**
  - ✓ Centralized integrity monitoring via System Health Center
  - ✓ Call home and remote service features
  - ✓ Fully redundant capability in all components for HA

- **Information Lifecycle Management (Release 1.1.1)**
  - ✓ Policy driven file placement, movement, migration and deletion of files over their entire lifetime
  - ✓ Storage tiering, support for SAS and SATA HDD's
  - ✓ Integrated TSM V6.1 HSM (space management) client for migration of inactive files to external TSM server

- **Disaster Recovery (Release 1.1.1)**
  - ✓ Asynchronous replication to another SONAS system

- **Public documentation / users manual / help center**

  http://publib.boulder.ibm.com/infocenter/sonasic/sonas1ic/index.jsp

\*NFS v4 soon, currently ACLs only

# SONAS / GPFS Architectural Scalability

**Max. Capacity (in 2010)**             14,4 PB

**Max. GPFS Design Capacity**         134,217,728 Yobibytes ($2^{107}$ Bytes)

**Max. Files   File systems  Snaps**    2 billion ($2^{31}$)   256   256

**Max. single File Size**            16 Exibytes ($2^{64}$ Bytes)

Lawrence Livermore National Lab



4096 Nodes

126GB/s *single file* write performance

# Disclaimer

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product data has been reviewed for accuracy as of the date of initial publication.  Product data is subject to change without notice.  This information could include technical inaccuracies or typographical errors.  IBM may make improvements and/or changes in the product(s) and/or program(s) at any time without notice.  Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The performance data contained herein was obtained in a controlled, isolated environment.  Actual results that may be obtained in other operating environments may vary significantly.  While IBM has reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere.  Customer experiences described herein are based upon information and opinions provided by the customer.  The same results may not be obtained by every user.

Reference in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.  Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used.  Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead.  It is the user's responsibility to evaluate and verify the operation on any non-IBM product, program or service.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED.  IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR INFRINGEMENT.  IBM shall have no responsibility to update this information.  IBM products are warranted according to the terms and conditions of the agreements (e.g. IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided.  IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein.

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 in connection with this publication 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.