Sun N1: Storage Virtualization and Oracle

Glenn Colaco

Performance Engineer

Sun Microsystems – Performance and Availability Engineering

September 23, 2003

We make the net work.

Background

- PAE works on database, CPU & systems, application server, & network performance.
- Close cooperation with Oracle.
- Provide feedback to:
  - ☞ Solaris[TM] Engineering
  - ☞ CPU / Systems & Compiler Engineering
  - ☞ ISVs (such as Oracle)

Overview

- Why is storage virtualization important?
- What work is being done in this space?
- Where is the industry going?

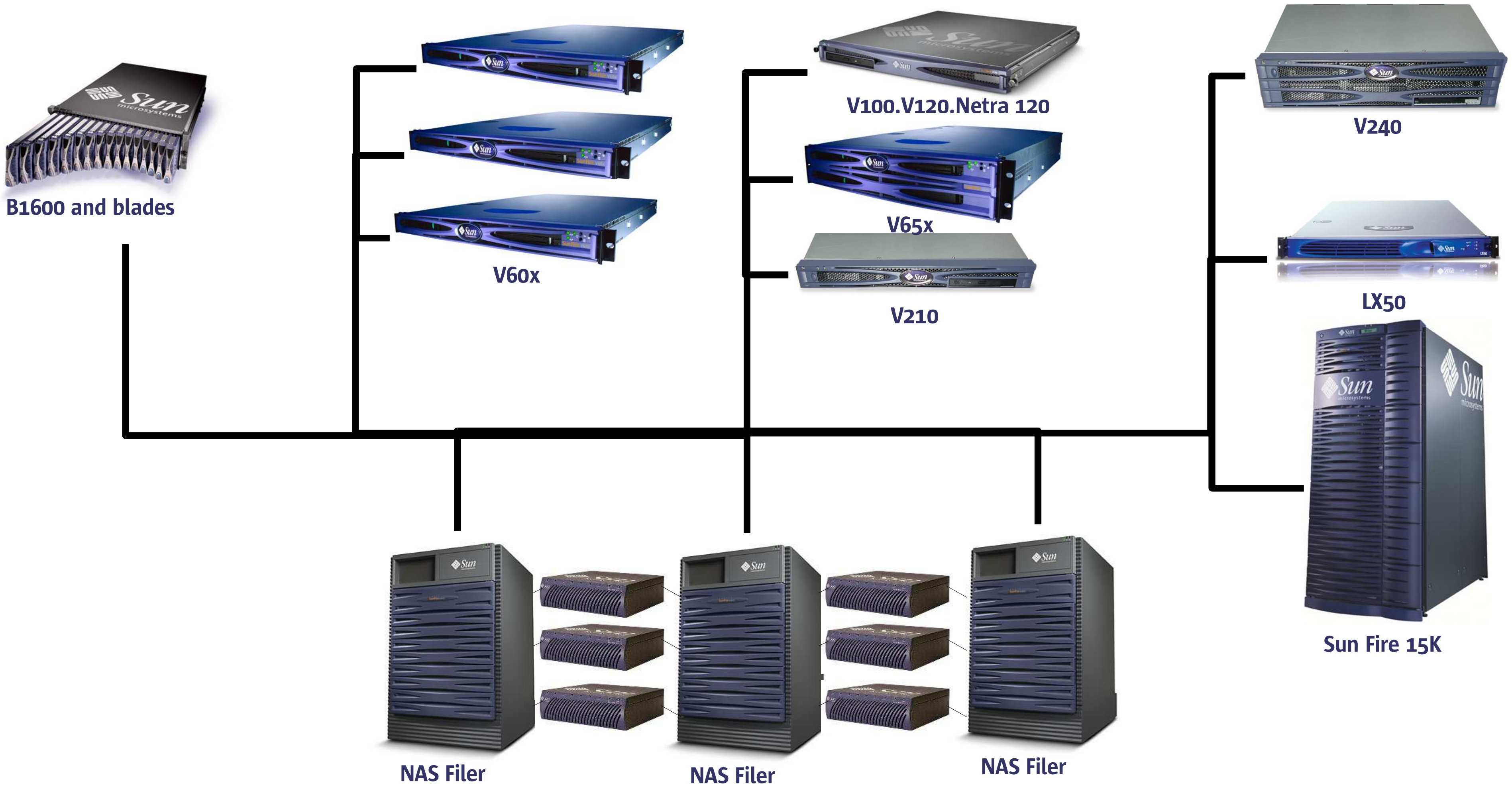
Overview

- Why is storage virtualization important?
  - ☞Overview of N1
  - ☞Overview of NAS
  - ☞Overview of Databases
  - ☞Why NAS?
  - ☞Common Perceptions and Misconceptions
  - ☞Storage/Network Interconnects
- What work is being done in this space?
- Where is the industry going?

Overview of N1

- Virtualization
  - ☞ Disassociate underlying system hardware and storage from application
  - ☞ Data is “available” anywhere on the network
  - ☞ Re-mapped onto any “compute element”
  - ☞ Grid computing
- N1 Database Model
  - ☞ Tier-3 is most complex to “virtualize” - compare to Tier 1
  - ☞ Provide support for RAC also
  - ☞ Must be high performing while also providing agility
  - ☞ NAS is critical for utility based DB deployment

An Enterprise IT Architecture



Overview of NAS

- Network Attached Storage (NAS)
  - ☞ Storage that is available via network, i.e. Ethernet
  - ☞ File or block based storage
  - ☞ NAS != SAN (Storage Area Network)

Overview of NAS

- History of NFS
- Network is the computer; data access should be available through network
- Based on open protocols as opposed to other “network” file systems at the time
- NFS created by Sun in 1984
- NFSv2 was released in 1985 and v3 in 1995
- File based access to data rather than block based access
- NFS is ubiquitous and available for most OS

Overview of NAS

- Database on NAS
  - Storage management is offloaded from the Database server
  - Decouples the storage management from Application and Database management
  - Database IO and storage requirements are drastically different than traditional NFS IO such as:
    - User's home directory
    - EDA market
    - Code development environments
- Where does NAS fit?
  - WAN as well as data center

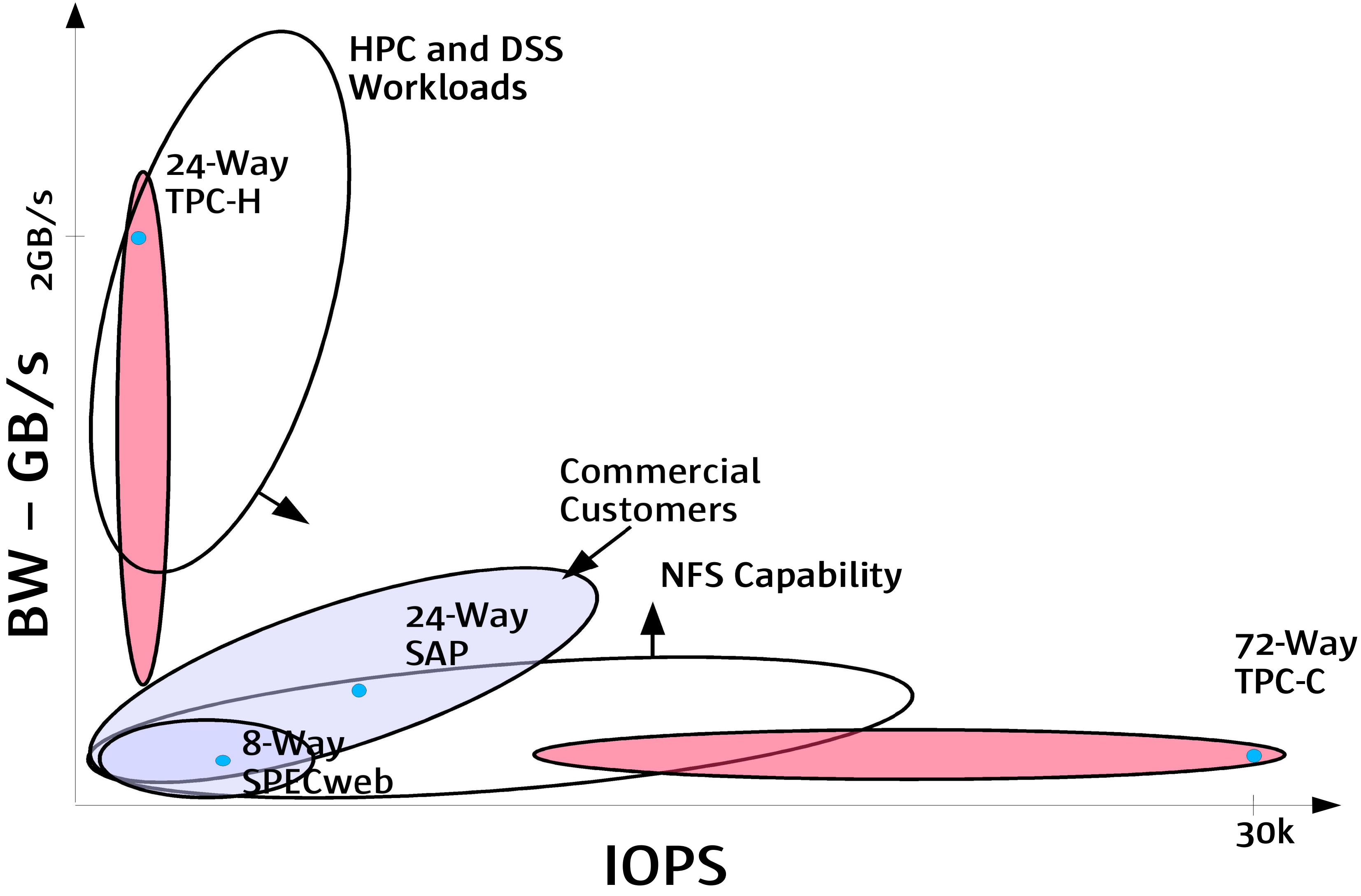
Overview of Databases

- Database consists of:
  - ☞Tables => Tablespaces => Datafiles => RAW/FS
  - ☞Datafiles as based on DB Blocks
  - ☞Asynchronous IO
- Online Transaction Processing (OLTP)
  - ☞DB Block Size = 2-8KB
  - ☞Random, parallel IO => High IO/s (on order of 10s of thousands), Low throughput (on order of 10s of MB/s)

Overview of Databases

- Decision Support System (DSS) aka. Data Warehousing/Data Mining
  - ☞ DB Block Size = 32KB
  - ☞ Large, sequential IO => Low IO/s (on order of 100s), High throughput (on order of GB/s)
- Customer workloads are a mix of OLTP and DSS

Benchmark Workloads in Today's Filesystem Landscape



Should I use RAW or Filesystems?

- Performance vs. manageability
  - ☞Single writer lock can be a problem
- Provide additional caching for DB
  - ☞32 bit vs. 64 bit applications
- Concurrent Direct I/O

Why NAS?

- Manageability
  - Storage
    - No more complexity of dealing with WWNs
    - Can build truly intelligent storage servers and provide extended file attributes
      - Provide additional Quality of Service (QOS) attributes and information
        - – storage server can now understand concept of milliseconds
        - “intelligent tablespaces”
    - NAS servers can understand Oracle file attributes and caching hints
      - No more BLACKBOX storage caching policies
      - Storage server doesn't have to “predict” what blocks should be in HW RAID cache
      - Rich protocol information tells HW RAID what Oracle datablocks should be cached

Why NAS?

- Manageability
  - ☞ Storage
    - – DAS (Direct Access Storage) storage is getting smarter and smarter, but only has intelligence of data blocks not files and extended attributes; blocks without context
      - ◆ c?t?d?s? has high response time, which tables are effected
    - – Can now manage Database by files which correlate to something intelligent such as tablespaces, etc.
  - ☞ Database
    - – Grow “tablespaces” with dealing with volume growth
    - – Can failover between nodes

Why NAS?

- Lowers overall cost of ownership
  - ☞ Commodity Hardware
- Simple Administration
  - ☞ Storage Consolidation
  - ☞ Eliminates need for client file system or volume manager
  - ☞ Fits well with organizational barriers
  - ☞ Appliance Model
- Blade Servers
  - ☞ Network based access only
  - ☞ Blade servers will have more compute power in the future

Why NAS?

- Alternatives
  - ☞Difficulties in DAS and block based storage
  - ☞iSCSI – just addressing transport and not the root problem

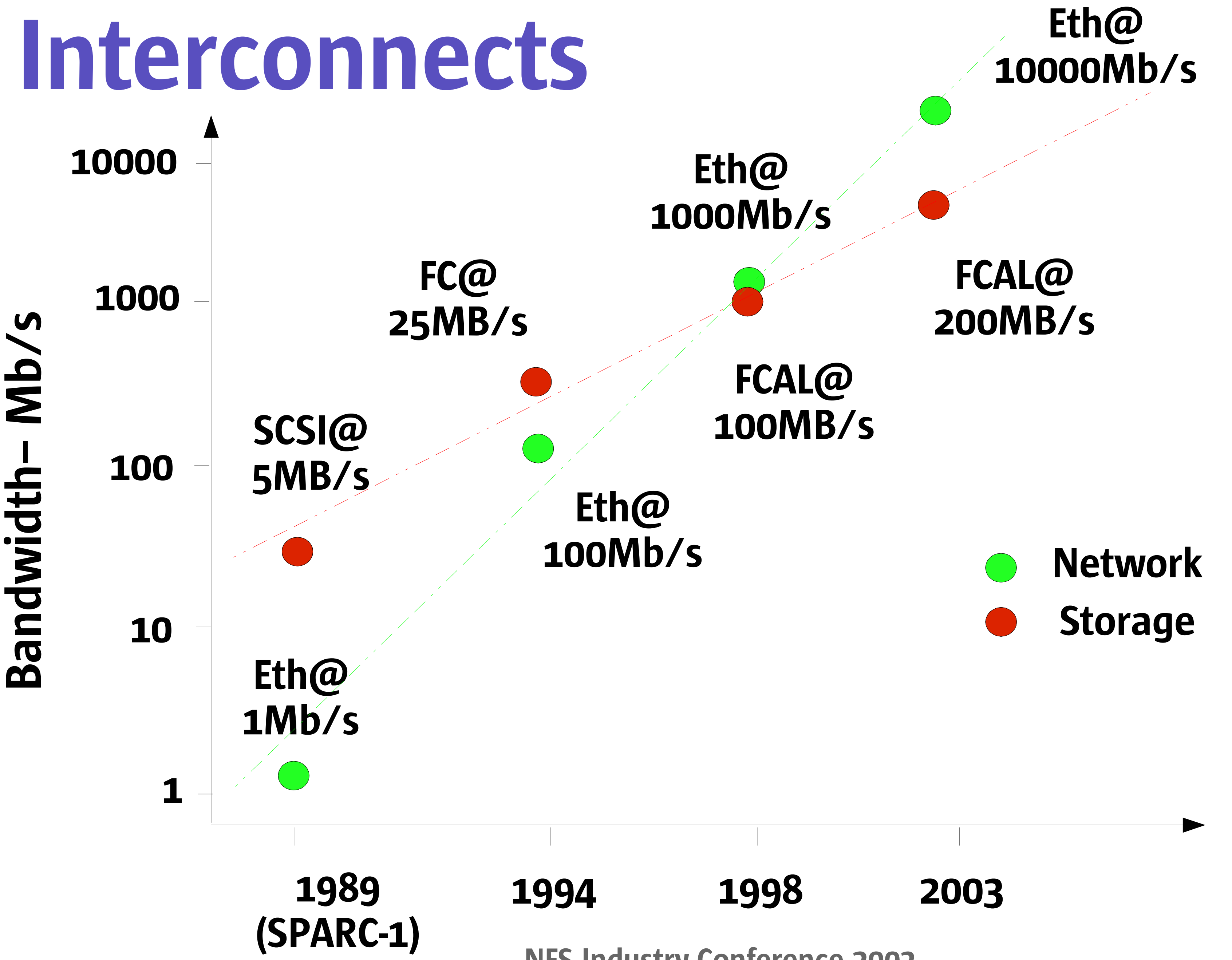
Why NAS?

- Alternatives
  - ☞ SAN vs. NAS
    - Block based IO is hard to manage
    - Tools not available
    - Data security is not as robust as NAS (i.e. IPSEC)
    - QOS not yet available
    - Reinventing the wheel
  - ☞ iSCSI vs. NAS
    - Block based IO is hard to manage
    - Utilizes same stack as NAS, but without the benefits of file based IO
  - ☞ Benefits of NAS
    - Transparent file access
    - Easier to grow storage pools
    - Easier to manage and backup; storage appliance

Common Perceptions and Misconceptions

- Performance
  - ☞ No one is running Databases on NAS
  - ☞ TCP/IP and network transport are the main problems
- Lack of Scaling
  - ☞ NAS won't meet high-end server requirements

Storage/Network Interconnects



Overview

- Why is storage virtualization important?
- What work is being done in this space?
  - ☞Project Background/Goals
  - ☞Performance Results
  - ☞Performance Enhancements in Solaris
- Where is the industry going?

Project Background/Goals

- Compare and improve Database NFS performance on NAS
- Contribute to industry direction
  - ☞ Infiniband vs. 10GE
- Sun's involvement
  - ☞ NFS over RDMA
- Parties involved
  - ☞ IETF
  - ☞ Key NFS vendors
  - ☞ Interface transport providers

Performance Results

- Compared both DAS connected and NAS connected storage
  - Database server used for DAS was exactly the same as the server for NAS
  - Direct connected Gigabit Ethernet – back to back
  - Using a well known OLTP workload, came within 15-20% of local UFS
    - OLTP workload generates on the order of 6x more IO than normal customer applications

Performance Enhancements in Solaris

- NFS Client:
  - ☞DirectIO 8KB write breakup
  - ☞Concurrent DirectIO
  - ☞Large IO transfers when using TCP
  - ☞RPC hashed wakeup mechanism
- Available in a Solaris 9U5 (12/03)

What about Network Attached Storage?

- OLTP vs. DSS is important
  - ☞OLTP is latency sensitive
  - ☞DSS is focused on bulk movement, so high throughput is needed
    - Jumbo Framing may be needed
- Verify both NAS client and server is optimized for Oracle performance

Overview

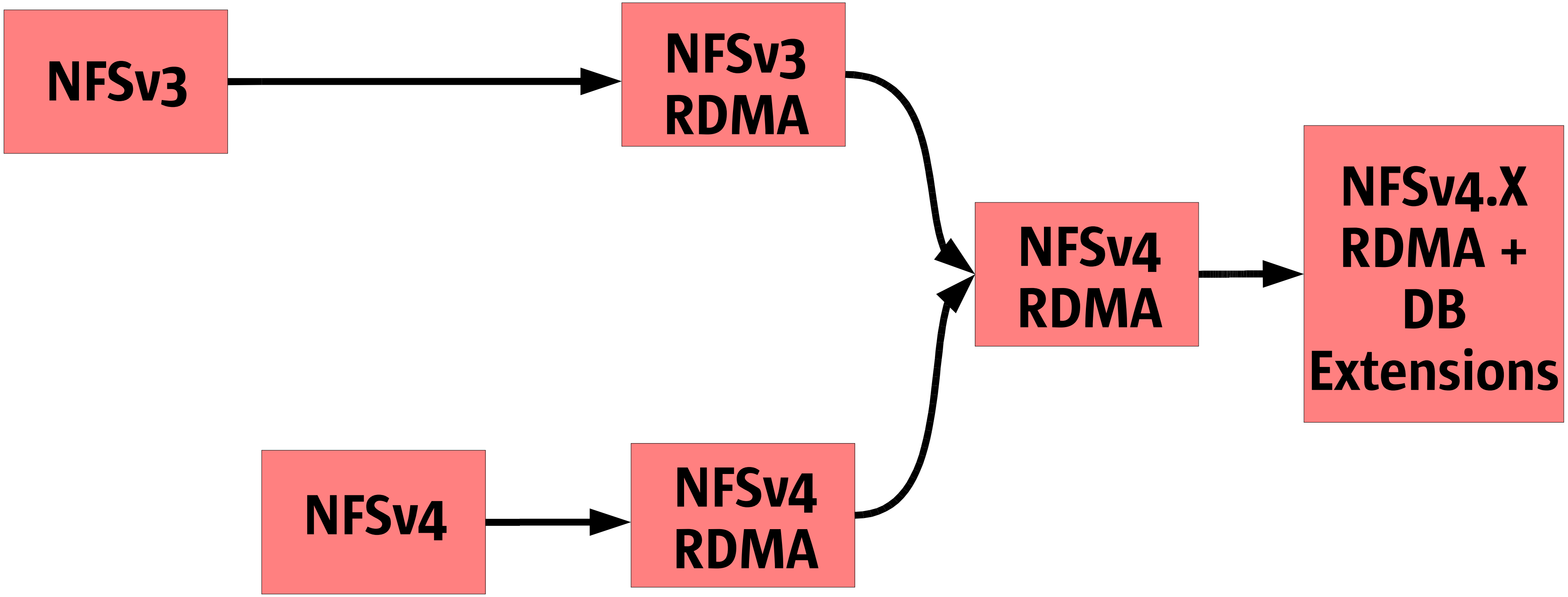
- Why is storage virtualization important?
- What work is being done in this space?
- Where is the industry going?
  - ☞Ongoing Industry Research
  - ☞Observations, Recommendations, & Speculations

Ongoing Industry Research

- RDMA – Remote Direct Memory Access
  - ☞ NFS over RDMA
  - ☞ IETF draft specification has been created and submitted
- TOE – TCP/IP Offload Engine
- Infiniband
- Evolution of NFSv4
  - ☞ NFS over RDMA
  - ☞ Database Extensions

Ongoing Industry Research

Where is the NFS protocol going?



Observations, Recommendations, & Speculations

- DirectIO
- TCP vs. UDP

Database NAS Performance: Whitepaper

- Paper will be available soon
- Please contact glenn.colaco@sun.com if you would like a copy when it is released

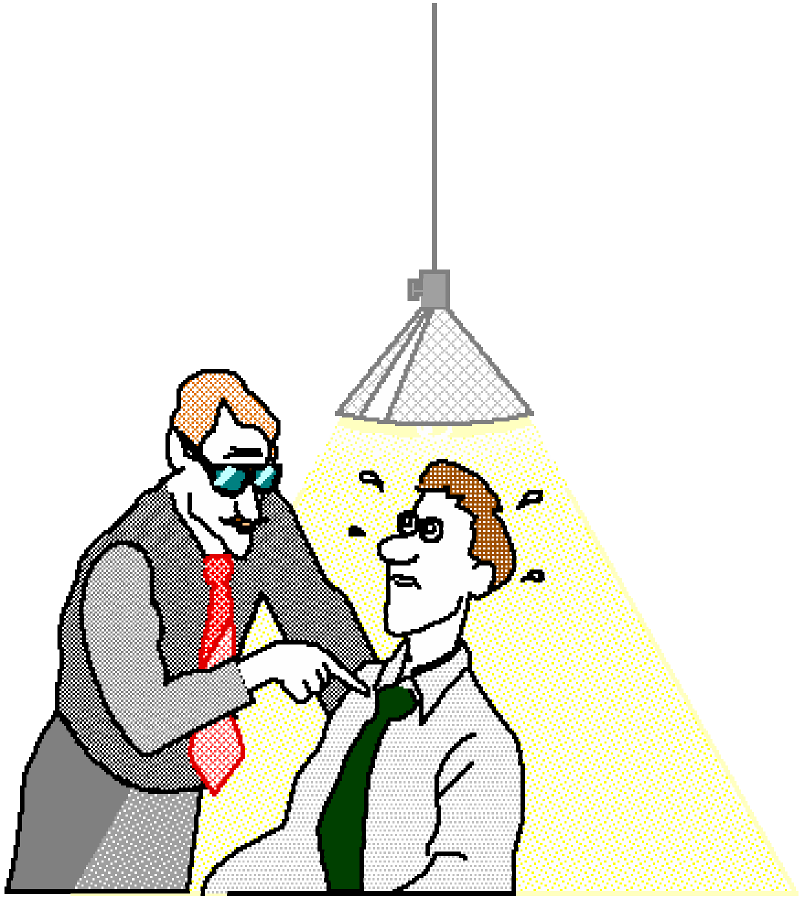
Conclusion

- Databases on NAS is not a bad idea
- New Solaris version will really help when dealing with OLTP
- With time, Databases on NAS will be a common practice

Feedback:

- Are you running DB over NAS?
- Who is interested in DB over NAS in the future?
- Would NAS simplify Database and storage management at your company?

Questions



Sun N1: Storage Virtualization and Oracle

Glenn Colaco

glenn.colaco@sun.com

September 23, 2003

Sun microsystems

We make the net work.