Analyzing Large-scale Network Boot Workload for Windows

Qiang Wang
Microsoft Corporation
Agenda

- Network Boot Basics
  - Windows iSCSI Diskless Boot
- Challenges of Large-scale Network Boot
- Workload Profiling and Visualization
  - Deployment and Boot of Diskless Windows Clients with Microsoft iSCSI Software Target
- Optimization for Boot
  - Hardware and Software Considerations
  - Microsoft iSCSI Software Target Scalability
Network Boot Basics

- Benefits of Booting from Network
  - Lower Capital and Operating Expenses
  - Better Manageability

- Network Boot Usages
  - Diskless Workstations
  - Thin Clients

- iSCSI Network Boot
  - Diskless Boot over Ethernet Fabric (RFC 4173)
  - Surfacing a Remote Virtual Hard Disk (VHD) as a Local SCSI Disk
iSCSI Bootstrapping

INT13

iSCSI Boot Initiator

Network Stack

iSCSI Boot Firmware Table

Pre-Boot

Windows

Boot Parameter Driver

Storage Stack

iSCSI Software Initiator

Network Stack

Standard Ethernet NIC
Challenges of Network Boot Scalability

- Server Service
  - High Throughput with Large Number of Active Sessions
- Storage I/O
  - Burst of Concurrent I/O Requests
  - Intensive Read Operations
- Network I/O
  - Load Balancing among Clients
- Management
  - Client Deployment
  - Client Servicing
Service Scalability - Threading

- Asynchronous Operations to Improve Processing Efficiency
  - Long I/O Operations
    - Network and Storage
  - I/O Completion Port
    - Queue for Completion Notifications

- Worker Thread Pool to Minimize Lifecycle Overhead
  - Serving Asynchronous Completion Notifications
  - Optimal Number of Worker Threads
    - Based on Number of Cores
    - Avoid High Unbalance

- Well-Designed Locking to Maximize Concurrent Processing
  - Lock-Free Data Structures and Algorithms
Service Scalability – Memory & CPU

- Memory Management
  - Memory Pool
    - Recovery from High Watermark
  - Disk Caching
    - Reducing Disk Physical Access
    - Leveraging OS Disk Caching
  - Tradeoff between Lower Memory Footprint and Higher Parallelism
    - Improving Lock Contention

- CPU Usage
  - CPU Cycles per Byte
    - Minimize Buffer Copy for Network and Storage I/O
  - Multi-Core Aware
    - Worker Threads to Core Mapping
    - Reducing Thread Context Switch Cost
  - Core Affinity
    - Network Receive-Side Scaling (RSS)
    - Session-to-Core Affinity
Scalability Tuning

- A Goal to Achieve Better than Linear Scalability
- Identifying Bottlenecks
  - Client Network Stack ↔ Server Network Stack ↔ Storage Service ↔ Storage Stack
  - Performance Counters
  - XPerf from Windows Performance Tools Kit
- Profile-Guided Improvement Iterations
  - Workload Profiling
  - End-to-End Analysis
  - Fully Understand Improvement and Degradation
Workload Profiling

- **Operation Types**
  - Read
  - Write
  - Management Task

- **Access Patterns**
  - Sequential
  - Random
  - Size
  - Burst
  - Locality

- **Network Load ≠ Storage Load**
  - Multicast Reducing Network Load
  - Disk Cache Reducing Storage Load
  - Multiple Disk I/O Operations in Response to a Single Network I/O Request
Boot Disk Workload

Disk Workload (Deploy2, all)

© 2010 Storage Developer Conference. © 2010 Microsoft Corporation. All Rights Reserved.
Understand Disk Workload Visualization

- **X-Axis**: Time
- **Y-Axis**: Logical Block Address (LBA)
- **Red Square** – Read
- **Blue Square** – Write
- **Gray Line** – Pseudo Disk Head Movement
  - Barely Visible for Sequential I/O
  - Easily Visible for Random I/O
- **Square Size** – I/O Size
- **LBA-to-File Mapping**
  - NTFS File Sector Information Utility
    - [http://support.microsoft.com/kb/253066](http://support.microsoft.com/kb/253066)

- **Data Collection**
  - XPerf
  - Custom Low-Overhead Tracing

- **Chart Generation**
  - XPerf Disk I/O Detail View
  - Your Favorite Graphing Tool

- **Visualization as Part of Analysis**
  - **Operation Type**
    - Read vs. Write
  - **Access Pattern**
    - Sequential vs. Random
    - Size
    - Burst
    - Locality
Boot Disk Workload Analysis

- **Raw Findings**
  - Dominance of Read Operations
  - Clear Disk Hot Spots
  - Identical I/O Pattern for all Clients

- **Addressing Read Scalability**
  - Decouple Read-Only and Read/Write Regions
    - Base OS Image Shared by All Clients for Read as a Base VHD
    - Each Client Has its Own Writable Region
    - Leveraging Differencing VHD Format
    - Diff VHD Stores Modified Data
  - Cache Disk Hot Spots
    - Leveraging NTFS Cache

- **Read/Write Size Histogram**
  - Bar chart showing read/write sizes from 1KB to 256KB
    - Frequencies range from 0 to 4000

Boot Disk Workload: Base vs. Diff

Disk Workload (Deploy2, base)

Disk Workload (Deploy2, diff)
Read Scalability Improvement

- 128 Clients Booting Simultaneously
  - Six Gigabit Ethernet Interfaces
  - One Logical Disk
- Green Line – % Disk Cache Hit
- Red Line – Disk Bytes Read/sec
- Six Dotted Lines – Network Bytes Sent/sec

- Disk Cache Hit Rate > 90%
- Aggregated Network Bytes Sent are 10x the mount of Disk Bytes Read
  - Most Network Bytes Sent Directly from Disk Cache
  - Remove Disk Read Bottleneck
Diskless Client Deployment

- Windows Deployment
  - Golden OS Image as Base
    - Sysprep’ed VHD
    - Used as Read-Only Parent VHD
    - Size for Windows 2008 R2 Server: Minimum 15GB
  - Differencing VHD for each Client
    - Read/Write
    - Modified Data Only
    - Typical Size after Deployment: 800MB
      - Grows as Needed by System
Deployment Disk Workload
Deployment Disk Workload Analysis

- Raw Findings
  - Similar Workload as Boot on Base VHD
  - Extensive Writes on Differencing VHD
  - Number of Differencing VHDs same as Number of Clients
    - Virtually any I/O Pattern on a Single Differencing VHD Becomes Random at System Level
- Addressing Write Scalability
  - Increase Number of Disk Spindles
  - Optimal Disk RAID Type
    - RAID 10 Performs Better than RAID 5 for Random Writes
  - Optimal Disk Controller Settings
    - Optimized for Write
      - Write Caching
Microsoft iSCSI Software Target Scalability

<table>
<thead>
<tr>
<th>Number of Clients</th>
<th>Deployment Time (min)</th>
<th>Boot Time (min)</th>
<th>Storage Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>30</td>
<td>4</td>
<td>20-Disk RAID 5</td>
</tr>
<tr>
<td>128</td>
<td>20</td>
<td>3</td>
<td>20-Disk RAID 10</td>
</tr>
<tr>
<td>256</td>
<td>58</td>
<td>6</td>
<td>20-Disk RAID 5</td>
</tr>
<tr>
<td>256</td>
<td>34</td>
<td>NA</td>
<td>20-Disk RAID 10</td>
</tr>
<tr>
<td>450</td>
<td>NA</td>
<td>9</td>
<td>20-Disk RAID 10</td>
</tr>
</tbody>
</table>

- Windows Storage Server 2008 R2
  - Single iSCSI Target Server
  - Twenty 10K RPM SAS Disks
  - Eight Gigabit Ethernet Interfaces
Summary

- iSCSI Network Boot
- Storage Server Scalability
  - Threading
  - Memory
  - CPU
- Workload Visualization and Analysis
  - Boot
  - Deployment
- Microsoft iSCSI Software Target Scalability
Appendix: Opportunities for Further Optimization

- What is the Silent Period in the Middle of a Deployment?
  - PnP Device Detection
  - Don’t Unnecessarily Generalize during Sysprep

- Which Files are the Hottest during Deployment?
  - Windows\System32\config\SYSTEM
    - Count: 3759 (12%)
    - Size: 4KB (88%)
  - $Mft
    - Count: 3710 (12%)
    - Size: 4KB (58%), 1KB (39%)
  - $LogFile
    - Count: 2373 (8%)
    - Size: 4KB (73%)