



Performance Benchmark IBM XIV Storage System

Whitepaper

Version: 1.3

Autoren: Bernd Patolla / Michel Centi

Datum: 11. August 2010

Klassifikation: **nicht klassifiziert**

In&Out AG IT Consulting & Engineering
Kilchbergsteig 13, CH-8038 Zürich, Phone +41 44 485 60 60
Fax +41 44 485 60 68, info@inout.ch, www.inout.ch



Management Summary

2009 lancierte IBM mit der XIV ein Storage System mit einer völlig neuen Storage Architektur.

Die IBM XIV ist aus Standard-Komponenten aufgebaut und basiert auf unabhängigen Modulen, von denen sechs Module Schnittstellen für die Storage-Anbindung bieten. Jedes Modul besitzt einen eigenen Controller, Cache-Speicher und 12 SATA-Disks. Die System-interne Kommunikation erfolgt über mehrfach redundante 1 Gbit / 10 Gbit-Ethernet-Verbindungen.

Durch den Einsatz von Standard-Komponenten bietet das IBM XIV Storage System eine sehr gute Performance. Dies auch in Kombination mit einer sehr einfachen Konfiguration und Administration. Das heisst, dass die Performance "out-of-the box" verfügbar ist, und nicht wie bei klassischen Storage-Systemen durch stetige aufwendige Konfigurationen sichergestellt werden muss. Da die XIV ganz ohne RAID Arrays und dem Platzieren von LUNs auskommt und die Verteilung der Daten automatisch über alle Disks im System erfolgt, ist das Storage System jederzeit optimal konfiguriert.

Wichtigste Erkenntnisse in Kürze

Sehr gute Performance "out of the box"

Das IBM XIV Storage System bietet bereits mit einem "out of the box" Standard Setup (1 LUN pro Server) eine sehr gute Performance. Dieses Setup zeigt durchwegs eine bessere Performance als ein konventionelles Setup eines Storage Systems.

Einsatz von Standardkomponenten

Die XIV ist ein modulares Storage System, welches Standard Industrie-Komponenten wie zum Beispiel SATA-Disks und Ethernet-Switches verwendet. Das Zusammenspiel dieser Komponenten wird durch eine neuartige Architektur und intelligente Algorithmen sichergestellt.

Verfügbarkeit

Durch die Spiegelung der Daten über die einzelnen Module eines XIV Storage Systems kann der Ausfall einer oder mehrerer Disks oder gar eines ganzen Moduls ohne Datenverlust und nur geringen Performance-Einbussen abgefangen werden.

Durch eine schnelle und automatische Neuverteilung der Daten bei Ausfall von einer Disk (1 TB Disk ca. 30 min.) wird die Chance eines Ausfalls einer zweiten Disk während der Neuverteilung stark minimiert.

Verwaltung

Die XIV überzeugt vor allem im Bereich der Verwaltung und Konfiguration. Muss bei einem konventionellen Speichersystem die Planung und Konfiguration ständig aktuell gehalten werden, macht dies die XIV automatisch und ist jederzeit optimal konfiguriert.

1 Einleitung

1.1 Motivation

IBM stellt mit dem XIV Storage System eine komplett neue Storage-Technologie vor. Die XIV stellt ein Grid-System dar, während herkömmliche Storage Systeme monolithische Ansätze verfolgen. IBM setzt in der XIV Industrie Standard-Komponenten ein, und positioniert das System durch die neuartige Architektur im Enterprise Segment.

1.2 Osys AG

Osys AG die KOMPETENZSCHMIEDE wurde 1978 gegründet. Als führender IBM Premier Business Partner realisieren die rund 40 IT-Spezialisten erfolgreich anspruchsvolle Projekte im Bereich "Dynamische Infrastrukturen" im modernen Data Center. Osys ist ein von IBM autorisiertes Systems- and Storage Solution Center (SSSC), XIV Storage Solution Center (XSSC) und verfügt über fundiertes und ausgewiesenes technisches Know-how zu allen wichtigen Server Plattformen und Storage Produkten von IBM. Ihre Kompetenz stellt Osys AG auch als VMware Premier Partner täglich in den Disziplinen Infrastruktur Virtualisierung, Desktop Virtualisierung und Business Continuity bei namhaften Kunden unter Beweis.

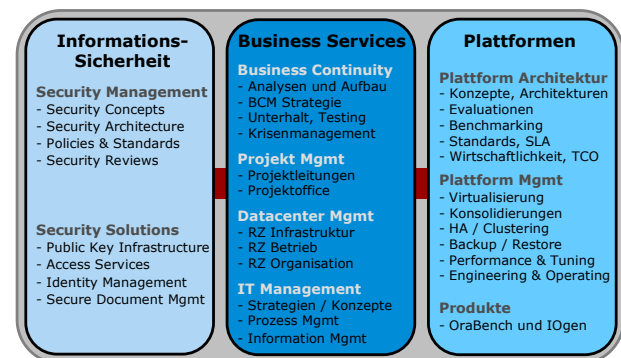
Weitere Informationen zur Firma sind auf dem Internet unter www.osys.ch zu finden.

1.3 In&Out AG

Kennzahlen

Die In&Out AG ist ein herstellernerutrales und unabhängiges Consulting und Engineering Unternehmen. In 2008 hat In&Out mit 34 Mitarbeitern 8.0 Mio. CHF Umsatz erzielt.

Dienstleistungen



Mit den drei Geschäftsfeldern Informations-Sicherheit, Business Services und Plattformen deckt In&Out eine breite Palette von Dienstleistungen ab.



Beim Betrieb von IT Plattformen ist es unerlässlich, dass die einzelnen Technologien optimal aufeinander abgestimmt sind

Die Fachkompetenz von In&Out deckt alle Komponenten ab.

Wir haben detaillierte Kenntnisse der Technologien und Produkte von führenden Herstellern, verfügen über jahrelange Praxiserfahrung im Zusammenhang mit Migrationen und Konsolidierungen, kennen die Produkte im täglichen Einsatz – und somit deren Vor- und Nachteile – und haben breite Benchmark-Erfahrungen.

Datenbanken
Server Management
OS
Server
SAN
Storage

Bei Plattform Projekten wie diesem ist eine übergreifende und durchgängige Sichtweise der verschiedenen Plattform-Layer vom Storage über Server bis hin zu Datenbanken oder Applikationen die besondere Stärke der In&Out AG.

Unsere Rolle

Im Speziellen setzt sich der Bereich Plattformen mit Optimierungen von High-End Systemen auseinander. In diesem Fall war die In&Out als externer Dienstleister verantwortlich für die Konzeption, Durchführung und Auswertung des Benchmarks.

Eingesetzte Tools

Für die Vermessung von Storage-Plattformen setzt In&Out ein eigens entwickeltes Produkt ein: IOgen™.

IOgen™

IOgen™ ist ein IO-Lastgenerator, welcher es ermöglicht, vordefinierte IO-Profilen auf unterschiedlichsten Plattformen automatisiert ablaufen zu lassen.

Die damit erzielten reproduzierbaren und vergleichbaren Messresultate können in nachfolgenden Tests höherer Layer (z.B. Oracle) als Grundlage für den IO-Durchsatz verwendet werden.

Folgende typischen Key-Performance Indikatoren (KPI) von Plattformen werden durch IOgen™ ermittelt:

- Geschwindigkeit: Anzahl IO Operationen pro Sekunde (IOPS)
- Durchsatz: Übertragene MB pro Sekunde
- IO-Servicezeit (SVT): Übertragungsdauer pro IO

2 Testszenarios

In beiden Szenarien standen insgesamt 1 TB Storage-Kapazität zur Verfügung. Beim Provisionieren dieses Speicherplatzes wurden zwei unterschiedliche Methoden gewählt.

Beide Setups durchliefen die gleichen Testreihen.

Konventioneller Storage Setup

In einem ersten Test wurde ein "klassischer" Ansatz gewählt. Auf dem Storage System wurden 20 LUNs à 51 GB erzeugt und dem Server zur Verfügung gestellt. Auf

dem Windows-Server wurden diese LUNs als dynamische Disks konfiguriert und mittels Striping ein logisches Volume erzeugt.

XIV Standard Storage Setup mit 1 LUN

In einem zweiten Szenario wurde ein Standard XIV Setup gewählt. Auf der XIV wurde dem Server eine 1 TB LUN zur Verfügung gestellt, welche im Windows-Betriebssystem direkt als Disk benutzt wurde.

3 Benchmark-Setup

3.1 Hardware

Server

- IBM X3950 M2
- Vier Intel X7460 CPUs à 2,6 GHz mit je sechs Cores
- 64 GB Memory
- Windows 2008 SP2
- Vier QLogic QLE2460 PCI Express HBA à 4 Gbit
- Windows Multipathing Treiber mit Round Robin

Während des Benchmarks wurde ein Windows 2008 Server als Basis benutzt. Der Server wurde mit vier HBAs der Firma QLogic à 4 Gbit ausgestattet und an ein redundantes SAN, bestehend aus je einem Switch, angeschlossen. Das Storage System war während der Benchmarks mit jeweils drei Verbindungen in jede Fabric angeschlossen.

IBM XIV Storage System

- IBM XIV Storage System (HW Release 2.0, SW 10.1.0.b)
- 6 Interface Module (erstes 1 CPU Model), 7 Data Module
- Cache: 104 GB RAM, davon sind ca. 65 GB dynamischer Lese/Schreib-Cache, gespiegelter Schreib-Cache, jedes Modul verwaltet seinen eigenen Cache, bei Ausfall eines Moduls fällt dessen Cache weg
- Kapazität: 156 TB (RAW), 1 TB / 7.2k RPM SATA Disks, 12 Disks pro Modul, insgesamt 156 Disks, 66 TB Nutzkapazität mit RAID X geschützt (jeder 1 MB / Block redundant vorhanden), 18 TB Spare Kapazität

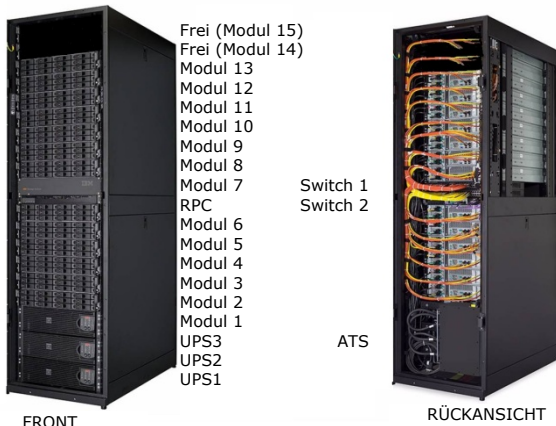


Abbildung 1 - XIV physikalischer Aufbau

Das IBM XIV Storage System ist ein selbstverwaltetes Grid aus 6 bis 15 Modulen. Einzigartige Funktionen wie der patentierte vollautomatische Verteil-Algorithmus halten die Daten jederzeit optimal verteilt, sei es nach Hardware-Veränderungen oder nach der Provisionierung von zusätzlicher Speicherkapazität. Die Verteilung der Daten auf bis zu 180 Festplatten, das Eliminieren des RAID 5 Write Penalty, die lineare Skalierung beim Hinzufügen von Modulen (Cache/CPU/Disk/IO-Controller) und die Minimierung der IO-Zugriffe (z.B. Snapshot mit Redirect on Write) erlauben den Einsatz von grossen SATA Platten im oberen Midrange-Segment.

Die Fibre Channel Anbindungen waren im Test bedingt durch die SAN Switch GBIC-SFPs auf 2 Gbit/s beschränkt.

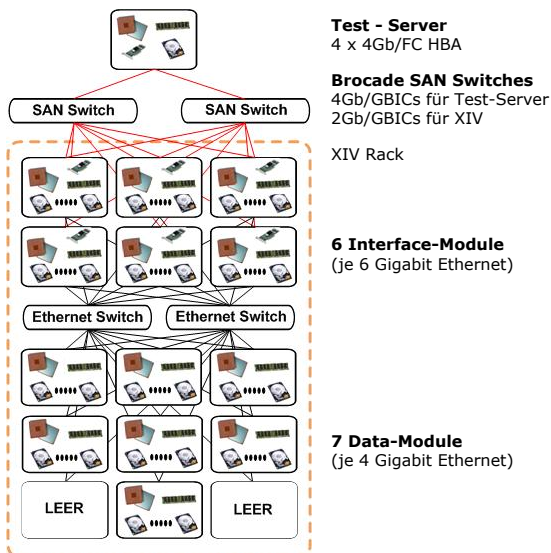


Abbildung 2 - Anbindung XIV - Test-Server

Intern kommunizieren die einzelnen Module mit je bis zu 6 parallelen Verbindungen über ein abgeschottetes redundantes Gbit-Ethernet (Load Balancing). Für die externe Anbindung besitzen gewisse Module zusätzlich iSCSI und 4 Gb FC-Schnittstellen (Interface-Module).

Der volle Leistungsumfang des IBM XIV Storage Systems wird durch die parallele Nutzung vieler Server und dem Einsatz von Zusatzfunktionen wie Snapshots erreicht.

3.2 Software

Lastgenerator IOgen™

- Version 2.3
- Random & Sequential IOs
- Blockgrößen: 8 KB & 1 MB
- Lesen & Schreiben

Lastprofile IOgen

Es wurde grundsätzlich zwischen Frontend- und Backend Tests unterschieden. In den Frontend Tests wird bis zum Cache des Storage Systems getestet, was durch die Beschränkung auf die ersten 10 MB der Disk(s) erreicht wird. Das bedeutet, dass das IO-Verhalten des Servers, die Konfiguration des Multipathing-Treibers, die HBAs und die SAN-Connectivity bis hin zum Caching des Storage Systems getestet werden.

Dagegen wird in den Backend-Tests bis auf die physikalischen Disks im Storage System getestet.

Sowohl im Frontend als auch im Backend werden die gleichen Tests durchlaufen:

- Zufälliges Lesen & Schreiben, Blockgrösse: 8 KB
- Sequentielles Lesen & Schreiben, Blockgrösse: 1 MB

Tests mit der Blockgrösse von 8 KB spiegeln dabei das Verhalten von Datenbanken und Applikationen wider. Die sequentiellen Tests mit Blockgrößen von 1 MB sind typisch für Backup-/Restore-Operationen oder Full Table Scans mit Multiblock-Read-IO in Oracle-Datenbanken.

4 Resultate

Die Ergebnisse dieser Tests sind in den folgenden zwei Kapiteln dargestellt. Die durchgezogene Linie stellt dabei die Lese-Performance, die gestrichelte Linie die Schreib-Performance dar. Die blauen Linien zeigen das klassische Setup (20 LUNs pro Server), die roten Linien das XIV Standard Setup (eine LUN pro Server).

Es werden zuerst die Ergebnisse der Frontend-Tests und anschliessend die Backend-Tests dargestellt.



4.1 Frontend-Tests

In den Frontend-Tests werden nur Zugriffe auf die ersten 10 MB einer jeden Disk durchgeführt. Somit kann das Storage System die Daten in den Cache laden und die Anfragen von dort aus schneller beantworten.

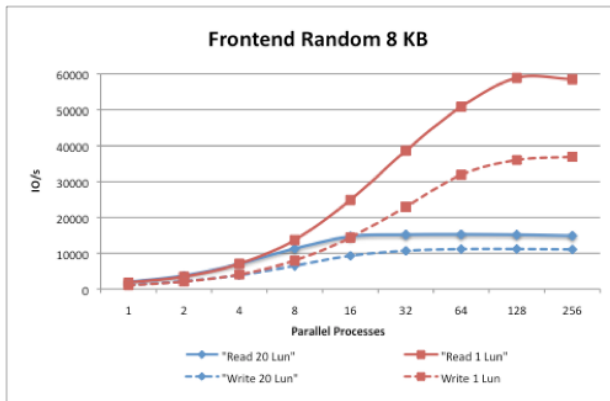


Abbildung 3 - Frontend Random 8K Blocksize

Auffällig ist, dass das klassische Setup ab mehr als 16 Prozessen keine Skalierung mehr zeigt. Dieses Setup erreicht den maximalen Wert von 15'200 IOs/sec lesend bei 32 Prozessen und bleibt auf diesem Niveau. Die Schreib-Performance fällt etwas tiefer aus, und liegt bei 11'200 IOs/sec. Sehr gut sind dagegen die Werte mit nur einer sichtbaren LUN am Server (XIV Standard Setup). Hier werden 60'000 IOs/sec lesend und 37'000 IOs/sec schreibend erreicht.

Durch die Begrenzung auf die ersten 10 MB einer jeden LUN wird bei dem Setup mit 20 LUNs eine Contention auf wenige physikalische Disks und einigen Modulen verursacht. Bei der XIV Standard Konfiguration mit einer LUN wird dies umgangen, da die XIV die Daten über mehrere Module und physikalische Disks verteilen kann.

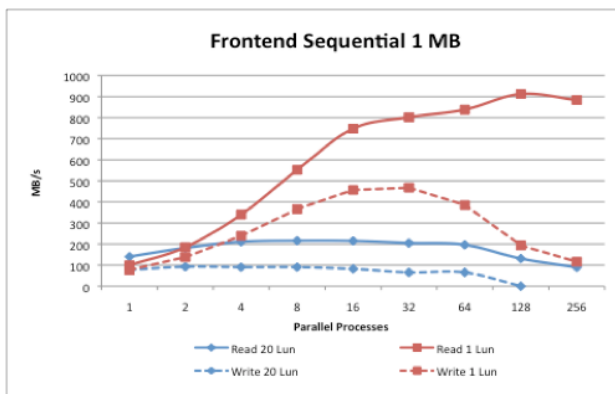


Abbildung 4 - Frontend Sequential 1MB Blocksize

Die sequentiellen Lese- und Schreibtests zeigen ein ähnliches Ergebnis. Das klassische Setup mit 20 LUNs erreicht schon sehr früh das eher geringe Maximum (210 MB/sec lesend, 90 MB/sec schreibend) und bleibt auf diesem Niveau. Dagegen skaliert das XIV Standard Setup bis auf 128 parallele Prozesse und erreicht dort 900 MB/sec lesend. Schreibend werden 470 MB/sec erreicht.

Die gemessenen Werte sind gut und erfüllen unsere Erwartungen in Bezug auf die Umgebung mit 4 Gbit HBAs und 2 Gbit Switches.

4.2 Backend-Tests

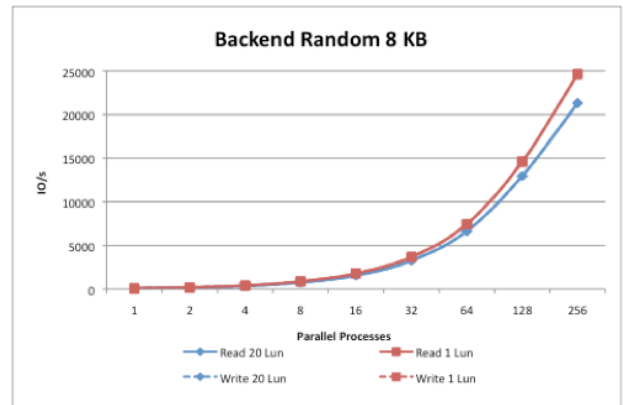


Abbildung 5 - Backend Random 8K Blocksize

Beide Setups (1 LUN und 20 LUNs) zeigen ähnliche Ergebnisse, da beide Setups die gleiche Anzahl physikalischer Disks im Backend verwenden. Die XIV liefert beim Lesen und Schreiben die gleiche Performance (die Kurven für Write-Operationen sind von den Read-Linien verdeckt).

Erst ab 64 parallelen Prozessen zeigt das XIV Standard Setup mit nur einer LUN eine leicht höhere Leistung wie das konventionelle Setup. So erreicht eine LUN knapp 25'000 IOs/sec, das klassische Setup bleibt leicht tiefer bei 21'000 IOs/sec.

Auch beim sequentiellen Lesen bietet das XIV Standard Setup die bessere Performance als das klassische Setup.

Die Unterschiede zwischen beiden Setups im Backend Random Read sind nicht so gross wie im Frontend Test da hier die gesamte Platte für Zugriffe benutzt wird. Es werden somit alle Module und physikalischen Disks belastet.

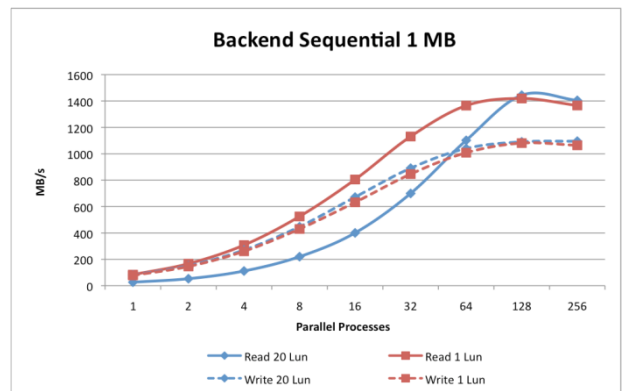


Abbildung 6 - Backend Sequential 1MB Blocksize

Beachtenswert ist bei diesem Test, dass das 1 LUN-Setup im mittleren Bereich fast doppelt so schnell liest (z.B. 16 parallele Prozesse: 800MB/sec vs. 400MB/sec),



bei hohen Parallelitäten dagegen beide wieder auf dem gleichen Niveau sind.

5 Zusammenfassung

Insgesamt zeigt das IBM XIV Storage System eine sehr gute Performance für ein Midrange-Storage System (wegen fehlender Mainframe z/OS Connectivity positioniert In&Out das System nicht im HighEnd Segment). Viele Unternehmen erreichen mit der Konfiguration ihrer HighEnd-Storage Systeme eine geringere Performance.

Beachtenswert ist dabei, dass das Setup mit einer einzigen LUN am Server eine weit bessere Performance bietet als ein herkömmliches klassisches Setup mit mehreren LUNs, welche durch das Betriebssystem noch gestriped werden müssen. Dies ermöglicht ein sehr performantes Setup mit einfachsten Mitteln "out of the box".

Dabei sollte beachtet werden, dass das XIV Storage System Standard-Komponenten wie SATA-Disks einsetzt. Durch den Einsatz einer internen Spiegelung der Datenblöcke ("RAID-X") über Modulgrenzen hinweg ist die Verfügbarkeit und Ausfallsicherheit gewährleistet.

Durch die Modularisierung sind relativ einfache Anpassungen an neue Technologien auch während des Betriebes möglich.

Die spezielle Architektur der XIV hat auch ihre Nachteile. So stehen aufgrund der internen Datenspiegelung und dem Bereithalten von Spare-Disks weniger als 50% der physikalischen Kapazität zur Verfügung.

6 Danksagung

An dieser Stelle möchten wir uns bei der Osys AG in Zürich-Oerlikon bedanken. Sie stellte die gesamte Test-Infrastruktur bestehend aus dem IBM Server X3950 über das SAN als auch das IBM XIV Storage System zur Verfügung.

Während der Tests unterstützten uns die Herren Felix Bürgler und Giacomo Chiapparini mit grossem Engagement.

Vielen Dank allen Beteiligten!

Der gesamte Inhalt dieses Dokuments ist geistiges Eigentum der In&Out AG und unterliegt dem Schutz des Urheberrechts.

Die vollständige oder teilweise Vervielfältigung, die elektronische oder mit anderen Mitteln erfolgte Verbreitung, die Modifikation oder die Benutzung für kommerzielle (insbesondere für Drittfirmen) oder öffentliche Zwecke bedarf der vorherigen ausdrücklichen Zustimmung durch die In&Out AG.

Die Verwendung des Whitepapers ist, unter Berücksichtigung des vorherigen Absatzes, nur als Ganzes mit unmodifiziertem Inhalt erlaubt.

7 Autoren



Bernd Patolla
Senior Platform Engineer
+41 79 831 19 61
bernd.patolla@inout.ch



Michel Centi
Senior Platform Engineer
+41 79 203 95 72
michel.centi@inout.ch



in&out

In&Out AG
Consulting & Engineering
Kilchbergsteig 13
8038 Zürich
+41 44 485 60 60
www.inout.ch



Giacomo Chiapparini
Senior Systems Engineer
+41 79 902 99 12
giacomo.chiapparini@osys.ch



Felix Bürgler
Systems Engineer
+41 79 297 11 27
felix.buergler@osys.ch



Osys AG
Hofwiesenstrasse 350
8050 Zürich-Oerlikon
+41 44 317 18 19
www.osys.ch