Filesysteme im heterogenen Zugriff

Frank Schmitz, Thomas Brandel
Forschungszentrum Karlsruhe
Institut for Scientific Computing (IWR)
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
Germany
www.CampusGrid.de

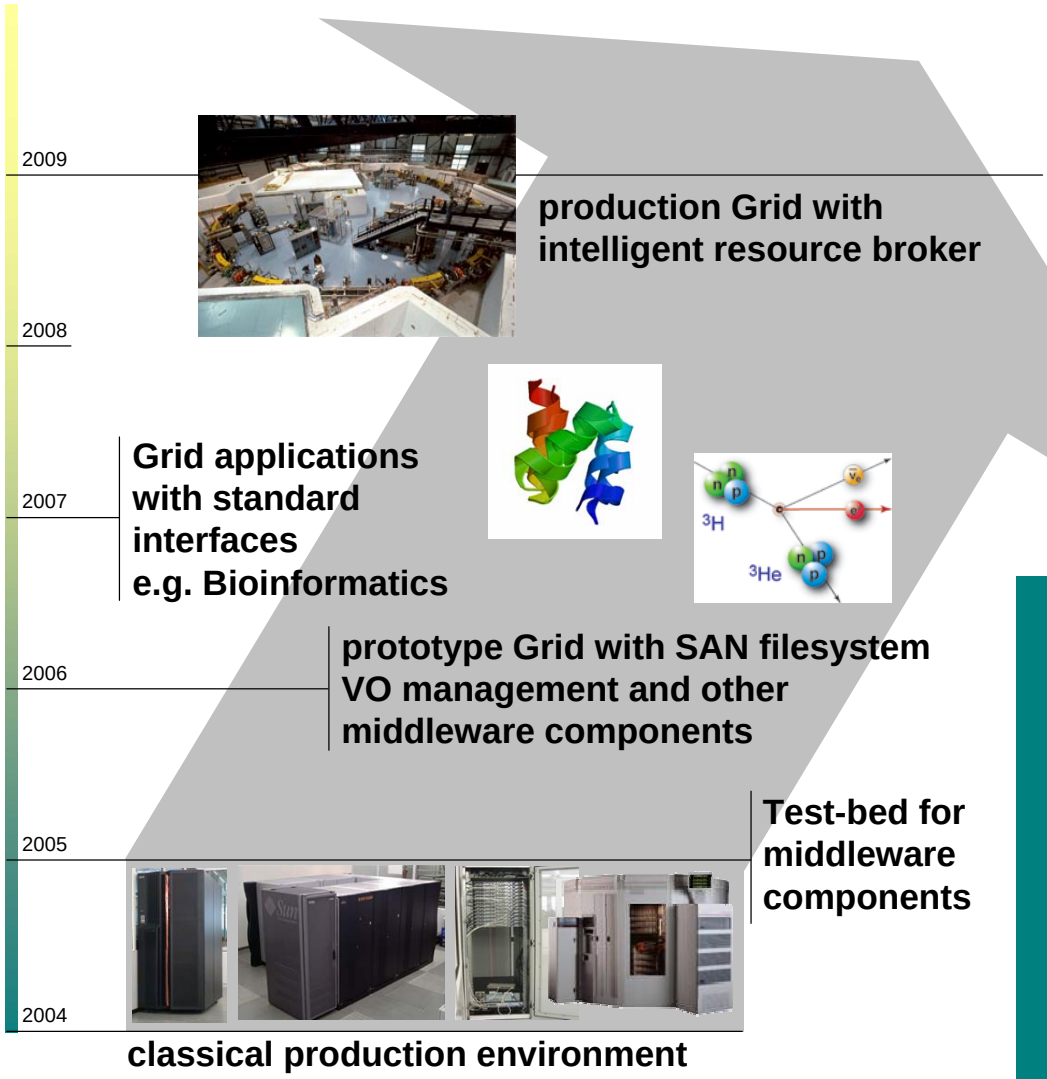
Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Motivation für das Projekt CampusGrid

- heterogenes Umfeld: Vektor-, SMP-, Cluster-, Blade-Systeme, SAN, NAS, Unix, Linux, Windows, Solaris, SuperUX, ….
- globaler Blick des Benutzers auf Daten, Hard- und Software
- nur ein Benutzermanagement
- ein globales Jobmanagement
- Metacomputing (MPI, ..), Echtzeitanwendungen
- Zugriff vom Arbeitsplatzrechner um Ergebnisse darzustellen
- globale Abrechnung
- nahtlose Integration in andere Konzepte und Projekte (gLite, LCG, D-Grid, Unicore, ..)

Ideen

- ein globales und schnelles Filesystem um allen Anforderungen genüge zu tun (StorNextFS, SAN-FS, SAM-QFS, NEC GFS, PVFS, Sistina GFS, CXFS, Celerra High-Road,…), Integration in ein langsames Grid-Filesystem
- Zugriffe über Infiniband, iSCSI, FC-SAN
- erste Schritte Ende 2004 → gLite als Middleware, aber …
- zweiter Versuch Mitte 2005 mit einer OGSA compliant Grid Middleware → Globus ToolKit 4 (aber nur ein Toolkit!)
- Ressource Broker (TORQUE, LSF, CONDOR, LoadLeveler…)
- Sicherheit → Kerberos 5 basierte Lösung
- Accounting → bisher Fehlanzeige (auf Lösungen von D-Grid und GridKa warten!)



Virtualization of a ***heterogeneous*** IT Environment for Scientific Computing

- R&D topics
  - Real-time data transfer protocols with efficient binary data
  - Grid file system
  - Global optimization (Resource broker)
- Several generations of test-beds

Produktionsumgebung und CampusGrid haben keine Schnittstellen!

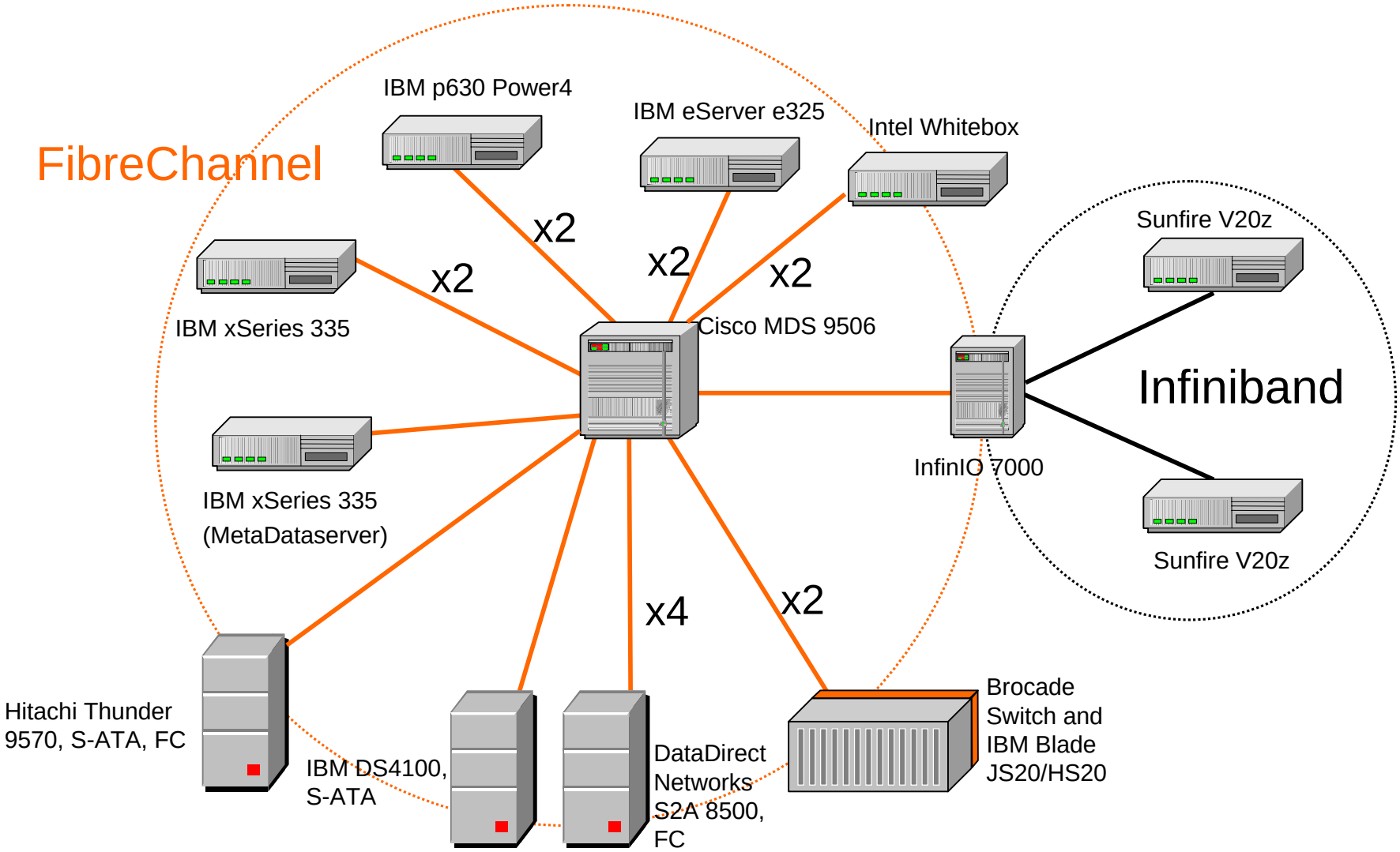
Ein Teil des CampusGrid Puzzels

Noch mehr Puzzleteile

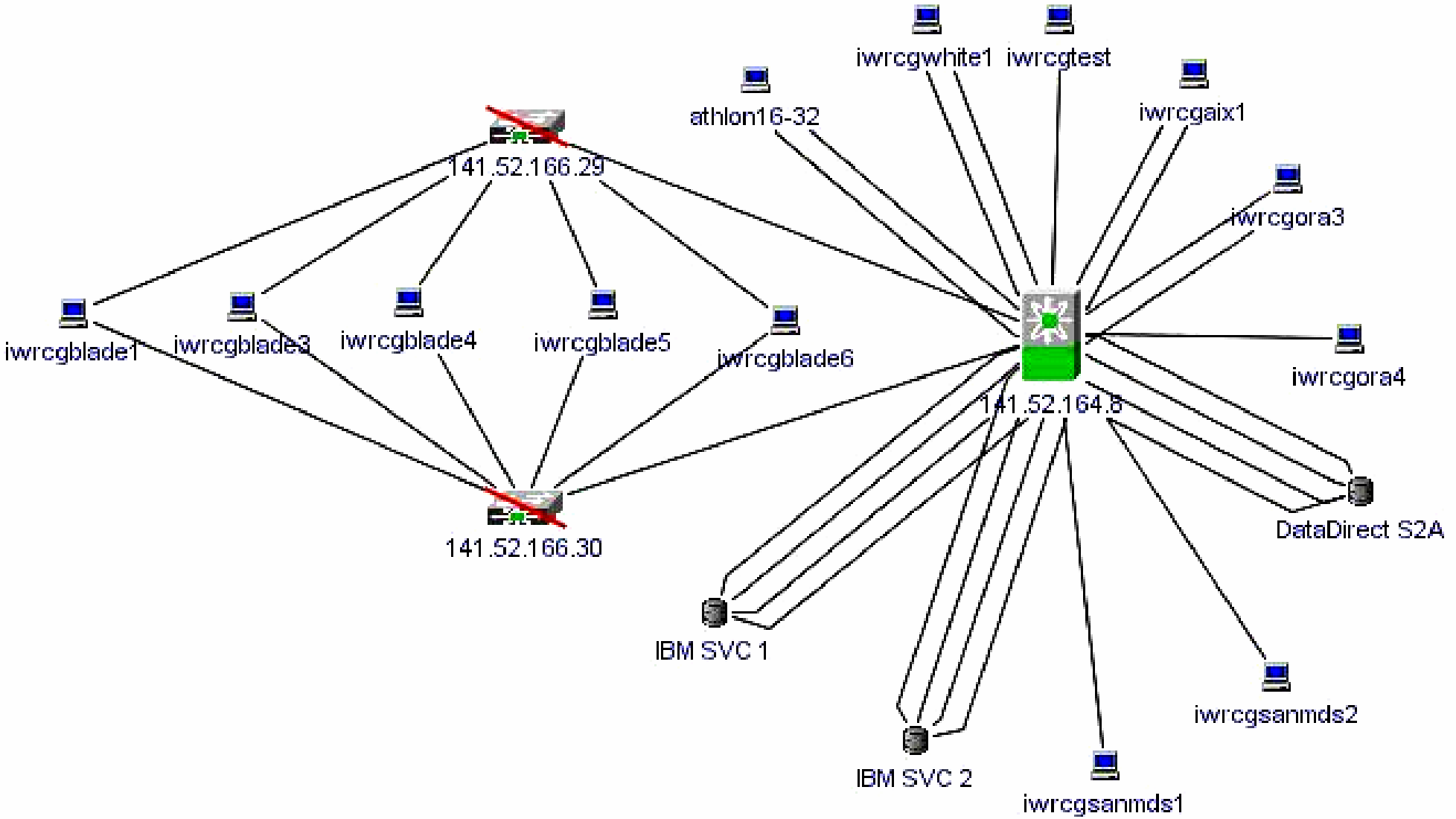
wassergekühltes Infiniband-Cluster mit 32 SUN V20z

Infinicon 9100, Infiniband Switch, MPI-Latency 4.0 µs zwischen den Knoten, 1.5 µs im Knoten

Das SAN Testbed



Noch mehr Details!



Die Hardwareumgebung

- Opteron dual prozessor system IBM e325, qlogic QLA2340 (athlon16-32)
- FibreChannel Storagesystem from DataDirect Networks (S2A 8500, 4.5 TB)
- Cisco MDS 9506 Director
- Intel Whitebox, Xeon, 2 MByte Cache
- IBM p630 (iwrcgaix1), AIX 5.2
- IBM x335 Dual Xeon
- Sun Vz20 Dual Opteron, InfiniBand (iwrcgtest2)
- InfinIO 7000
- IBM Bladecenter with JS20 and HS20
- IBM DS4100 (FastT100), S-ATA disks
- Hitachi Thunder 9570 SATA+FC (3.5 TB)

Die getesteten Filesystemlösungen (1)

ADIC StorNextFS V 2.4.1

Hardwarevoraussetzungen (MDS): min. 512 MB RAM, HBA(s)
Unterstützte MDS- und Clientplattformen:

- AIX 5.2 (64bit)
- IRIX 6.5.19, 22f und 6.5.19, 22 und 23m
- Solaris 8 und 9
- RedHat EL 3 Kernel 2.4.21-15.0.x-EL(smp)
- SuSE Linux Enterprise Server Kernel K_delft-2.4.19-304 bzw. K_smp-2.4.19-304
- Windows 2000 SP4, Windows NT SP6a, Windows XP SP2, Windows 2003 Server

Die getesteten Filesystemlösungen (2)

Installation MDS: unter Linux Einspielen von 2 RPMs (Client + MDS), Konfiguration über Textfiles
MDS installiert auf: IBM xSeries x335, RedHat AS 3.0, Kernel 2.4.21-4.ELsmp
Unterstützte Clientplattformen: wie Server
Sehr einfache Clientinstallation:
Paket einspielen, kurze Anpassung von Textfiles (Unix) bzw. GUI (Windows)
Client installiert auf:
IBM Blade HS20 (RH EL 3.0 U4 und Windows 2000 Server, dual Xeon), IBM x335 (RH EL 3.0 U4, dual Xeon), Intel Whitebox (RH EL 3.0 U4, dual Xeon EM64T), IBM e325 (RH EL 3.0 U4, dual Opteron), Sunfire V20z (RH EL 3.0 U4, dual Opteron, SAN-Anbindung über Infiniband-FC-Bridge), IBM p630 (AIX 5.2, 4x Power4)

Die getesteten Filesystemlösungen (3)

Pro:
- Hohe Heterogenität (Client und MDS)
- Einfache Installation von Client und Server
- Multipathing als Funktionalität im Filesystem
- Viele Kernelvarianten bei Linux sind möglich
- Gute Interoperabilität zwischen Unix und Windows (Groß-Kleinschreibung)

Kontra:
- In v2.4.1 nur 2 Metadatenserver möglich, kein Lastausgleich
- Nach Löschen von Devices auf p630 (rmdev) muss MDS neu gestartet werden

Zukunft
- mehr als zwei Metadatenserver, kein Lastausgleich
- Integration ins Active Directory
- 64-bit Linux

Die getesteten Filesystemlösungen (4)

IBM SAN-Filesystem V2.2.1

Hardwarevoraussetzungen (MDS):

- Unterstützt: je 2x IBM x345 oder x346
- Alternativ notwendig: 2 Server mit jeweils min. 4GB RAM, IBM RSA2-Adapter, 2 HBAs, 2 Prozessoren

Unterstützte MDS-Plattform: ausschließlich SuSE Linux Enterprise Server 8 SP 3

Installation MDS: Einspielen einer Reihe von RPMs, Konfigurationsänderungen am SLES 8 notwendig (kein X-Server, etc.), Konfiguration der Server über Konfigurationsskripte und nach der Grundkonfiguration über Weboberfläche.

MDS installiert auf: 2x IBM xSeries x345

Die getesteten Filesystemlösungen (5)

Unterstützte Clientplattformen:

- AIX 5.1 (32bit)
- AIX 5.2 (32bit und 64bit)
- AIX 5.3 (32bit und 64bit)
- Solaris 9 (64bit)
- RHEL 3 (nur x86, ausschließlich Kernel 2.4.21-15.0.3)
- SLES 8 (nur x86, ausschließlich Kernel 2.4.21-231)
- Windows 2000 (Advanced) Server SP4, Windows Server 2003 Enterprise und Standard Edition

Installation Client: Software installieren, einfaches Konfigurationsskript (Unix) bzw. GUI (Windows), Anbindung des Clients über Weboberfläche des Servers. Da unter Linux die unterstützten Kernel nicht installiert waren, mussten per Hand Änderungen am Installationsskript bzw. am System durchgeführt werden, um den Client zu installieren.

Die getesteten Filesystemlösungen (6)

Clients installiert auf: IBM Blade HS20 (RH EL 3.0 U4 und Windows 2000 Server, dual Xeon), Blade JS20 (AIX 5.2 und AIX 5.3, dual PowerPC), Intel Whitebox (RH EL 3.0 U4, dual Xeon EM64T), IBM p630 (AIX 5.2, 4x Power4)

Pro:

- Abgesehen von Linux (Kernel muss genau passen!) relativ gute Heterogenität bei den Clients
- Einfache Client-Installation und Anbindung an den MDS
- Sehr gute und intuitiv bedienbare GUI
- Viele MDS möglich

Kontra:

- Serverinstallation sehr komplex
- Authentifizierungsinformationen der Clients liegen im SAN, beim Neu-Aufsetzen des Filesystems müssen per Hand Dateien auf den MDS gelöscht werden, um die Authentifizierung neu aufzubauen. Bei Tests etwas nervig!

Die getesteten Filesystemlösungen (7)

- Für den „System Pool“ (Metadaten-LUNs) muss zwingend entweder IBM-Storage oder über SVC virtualisierter Storage eingesetzt werden.
- Unter Linux nur auf sehr speziellen Kernelversionen regulär installierbar (genau definierter Patchlevel)
- Kein automatisches Multipathing möglich (nur über Hersteller-Software des Plattensystems)
- Keine saubere Unterscheidung bei Groß/Kleinschreibung unter Windows
- kein Lastausgleich
- Spezielle Hardware für MDS erforderlich (RSA II)

Die getesteten Filesystemlösungen (8)

SGI CXFS V3.2.5 (AIX), V3.2.6 (Linux), V3.2.2 (Windows)

Hardwarevoraussetzungen (MDS): SGI-Hardware erforderlich
MDS-Plattformen:

- SGI IRIX 6.5.24 bzw. SGI Linux Kernel 2.4.21-27.0.4.EL.sgi10smp mit SGI ProPack

Installation Server: Wurde von SGI-Mitarbeiter durchgeführt
Server installiert auf: SGI Altix (Linux), ein Riesenhobel!
Clientplattformen:

- AIX 5.1 und AIX 5.2 (64bit) auf IBM pSeries
- RedHat Linux 9 bzw. RedHat EL 3
- Mac OS X 10.3.5
- Solaris 8 und 9
- Windows 2000 und XP

Die getesteten Filesystemlösungen (9)

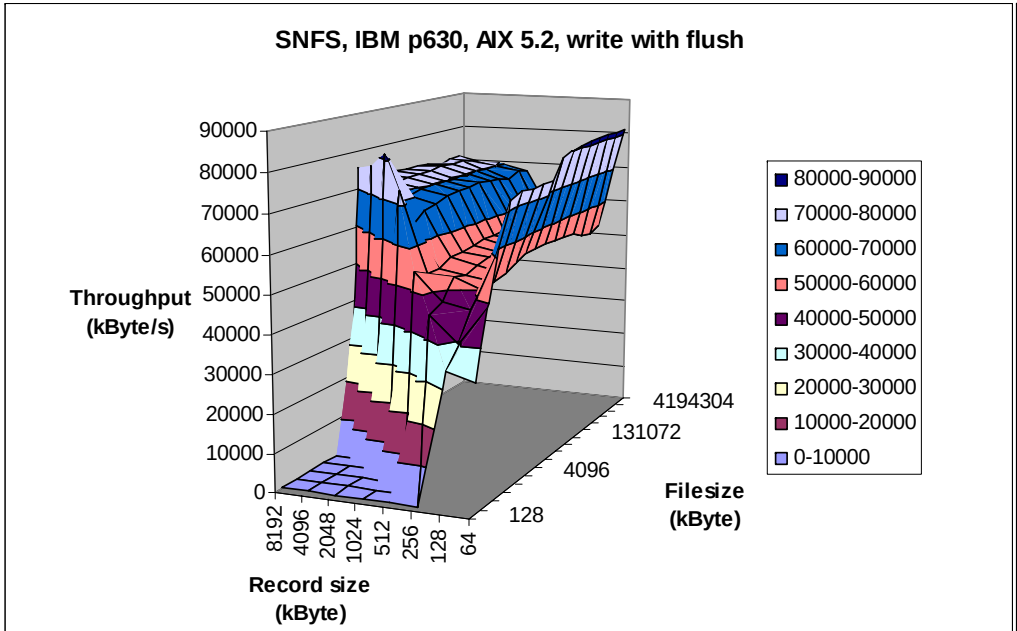
Installation Client: Unter Linux Einspielen einer Reihe von RPMs (u.a. eigener Kernel), unter AIX Installation eines Pakets, unter Windows graphische Installation. Danach muss ein Lizenzschlüssel installiert werden. Dann: Start des Filesystems, Einbinden des Clients über GUI auf Server.
Clients installiert auf: IBM Blade HS20 (Windows 2000 Server, dual Xeon), IBM p630 (AIX 5.2, 4x Power4), Intel Whitebox (RH EL 3.0 U4 mit SGI-Kernel, dual Xeon EM64T), Sunfire V20z (RH EL 3.0 U4 mit SGI-Kernel, dual Opteron, SAN-Anbindung über Infiniband-FC-Bridge)
Pro:
• Abgesehen von Linux relativ gute Heterogenität bei den Clients
• Einfache Anbindung der Clients über GUI, relativ einfache Installation der Client-Software
• Einfache Administration des Servers über GUI

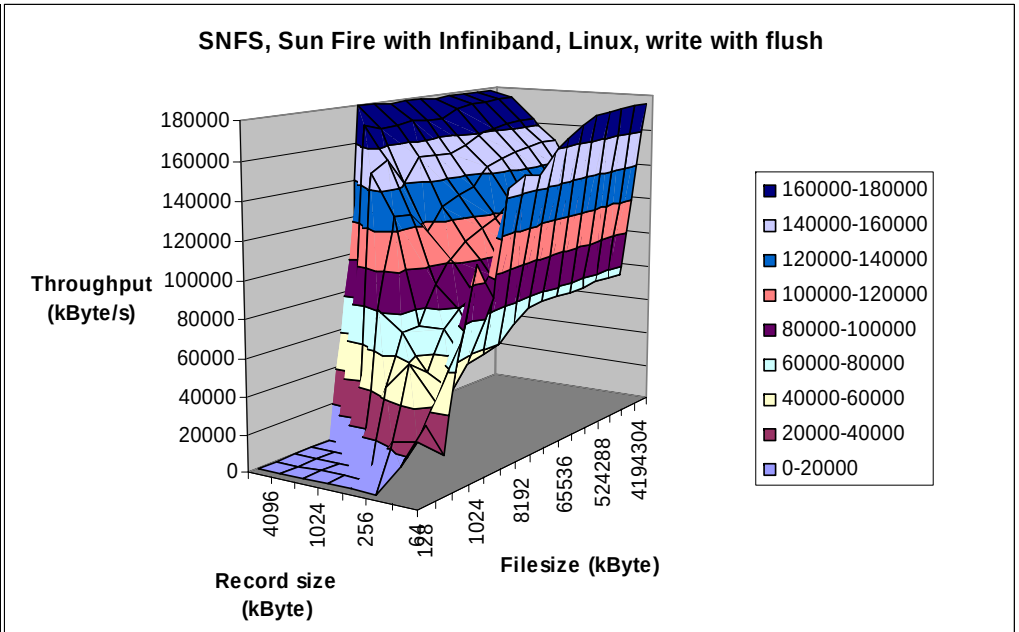
Die getesteten Filesystemlösungen (10)

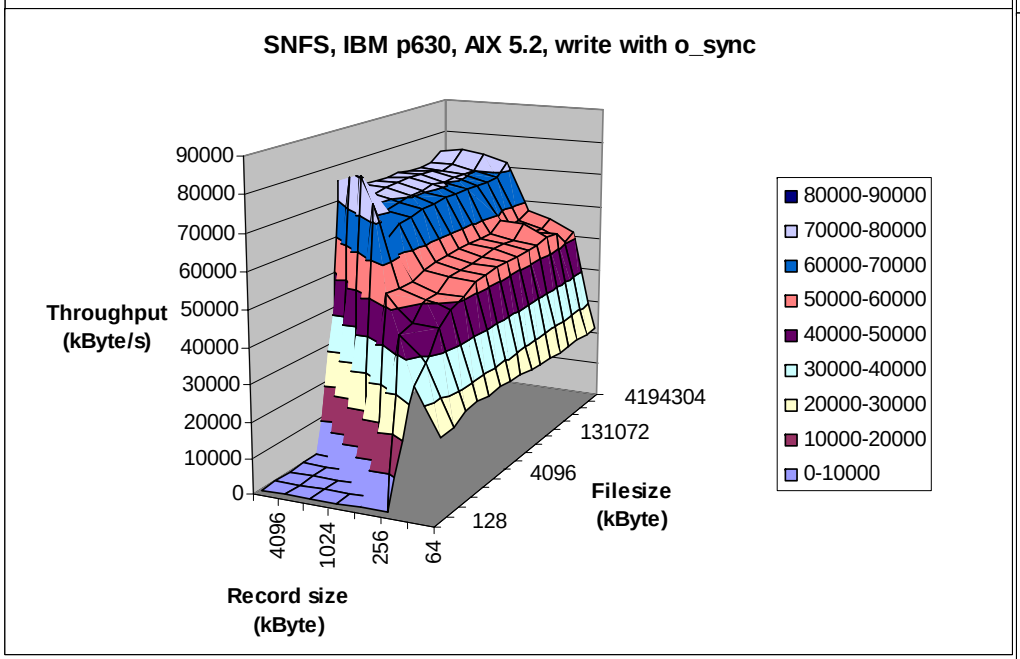
Kontra:

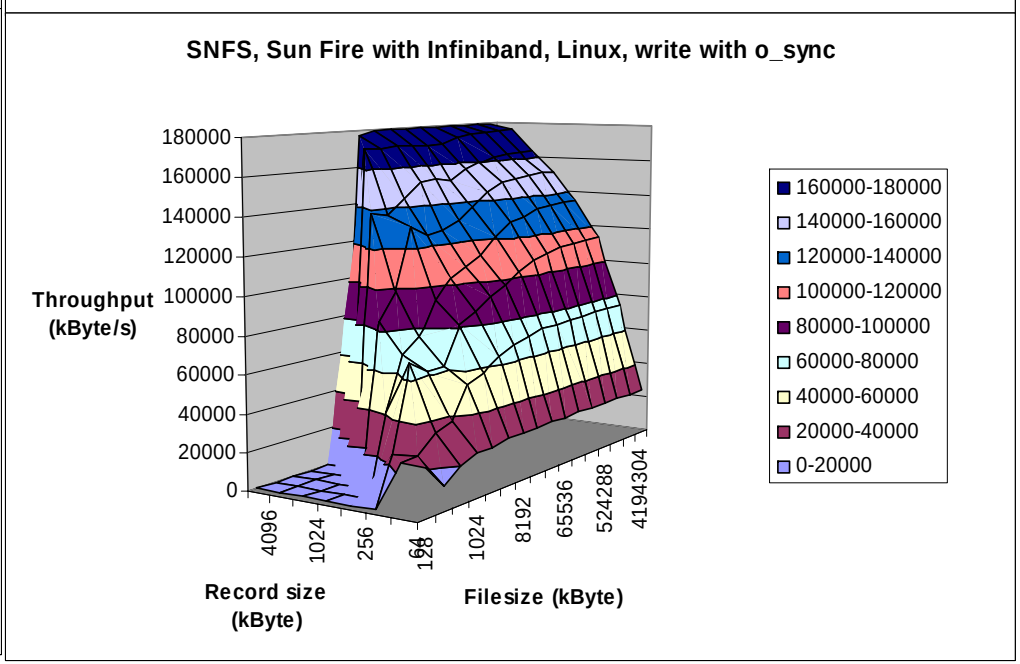
- Spezielle SGI-Hardware erforderlich
- Clients und Server müssen in demselben Subnetz liegen
- Spezieller SGI-Kernel unter Linux erforderlich
- Kein Multipathing über das Filesystem
- Getestete Version war auf PowerPC (Blade JS20) unter AIX nicht lauffähig
- Lizenzschlüssel an Hardware gebunden

ADIC SNFS Single Path

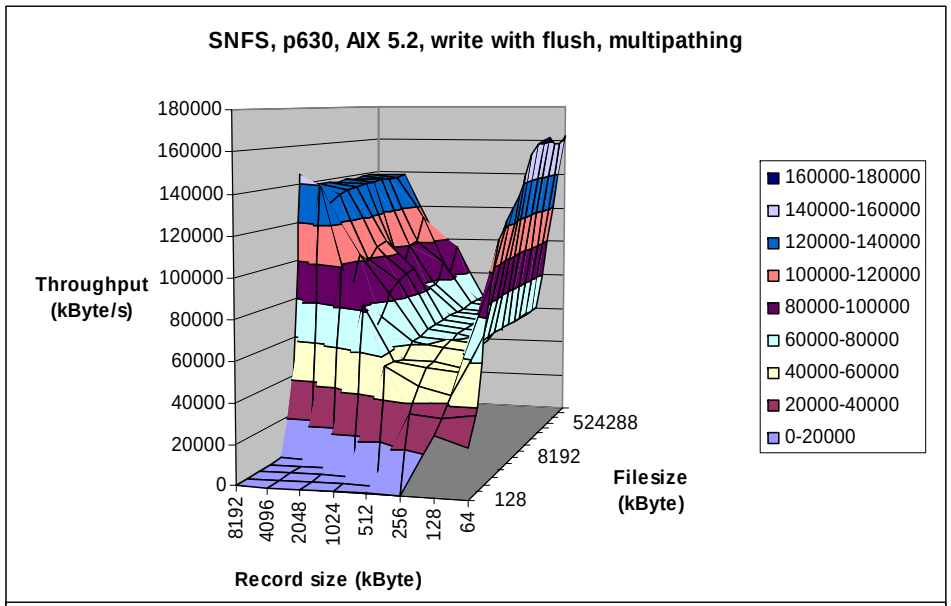


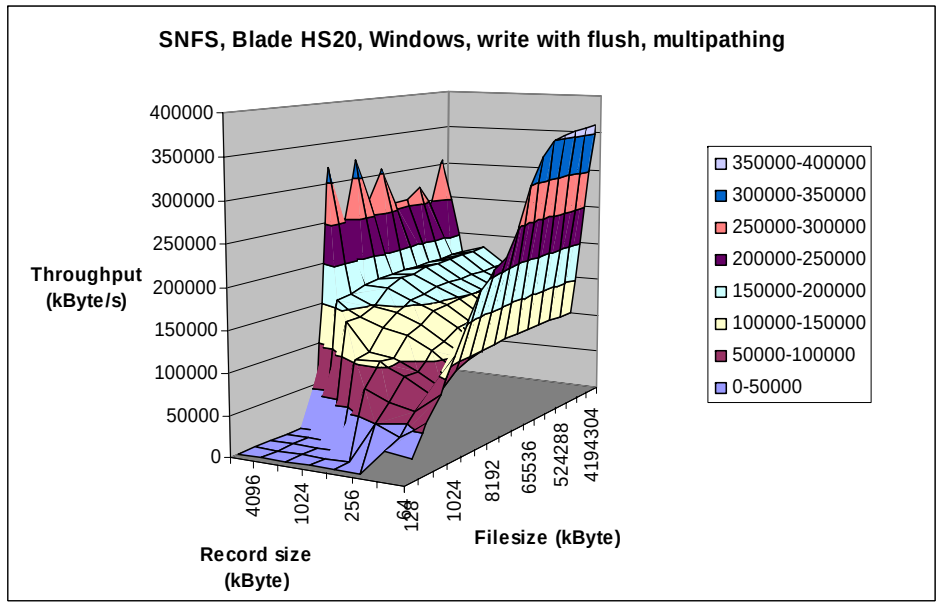


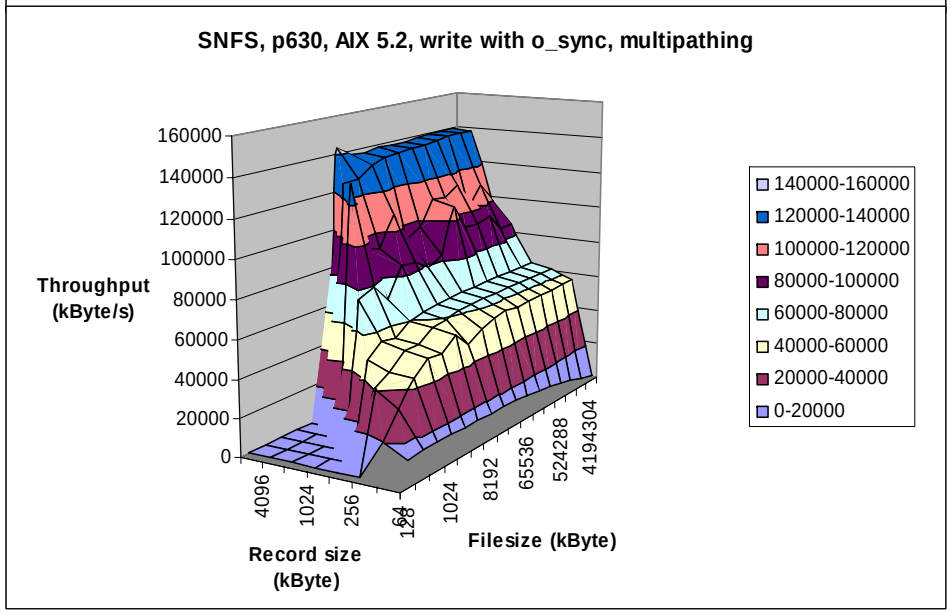




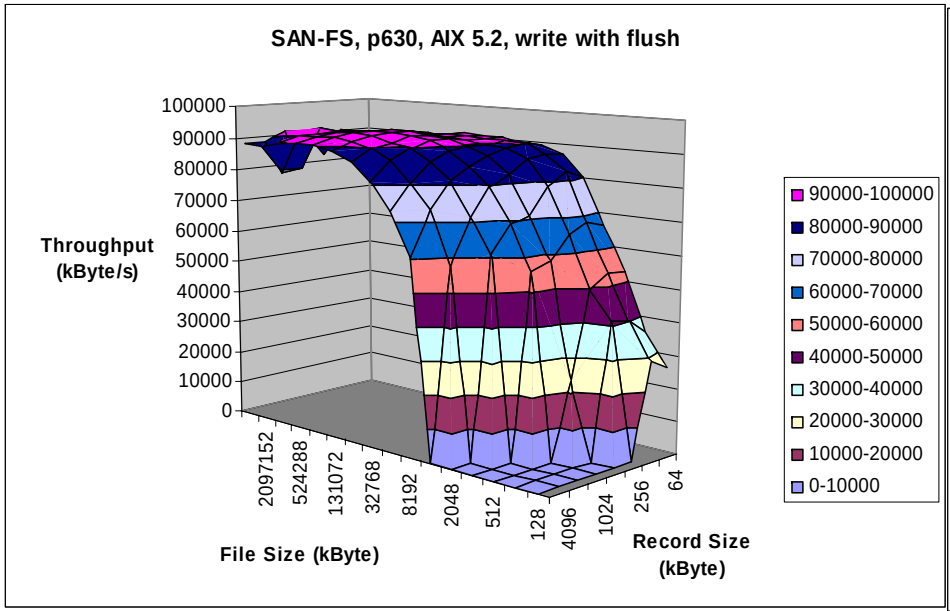
ADIC SNFS Dual Path

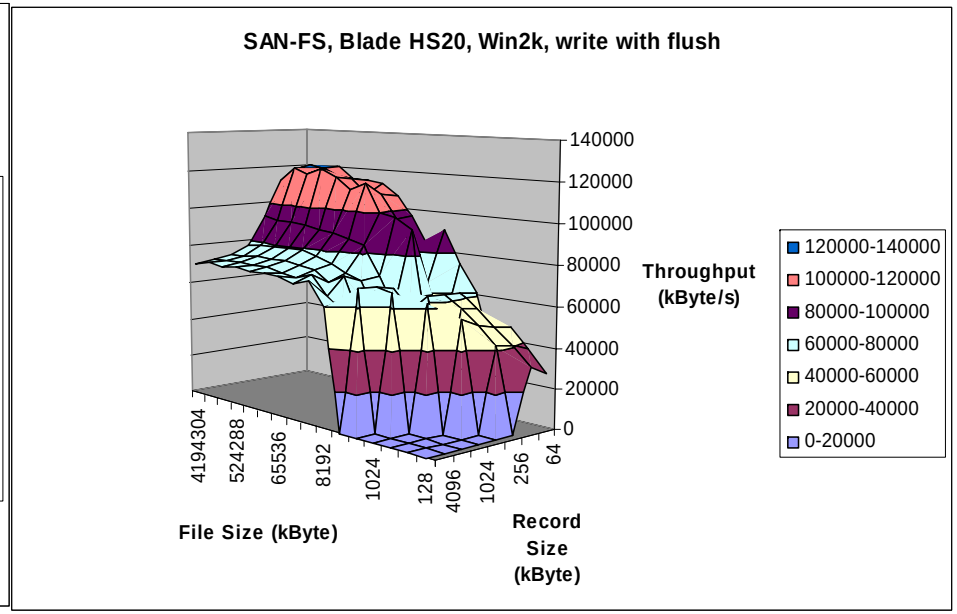


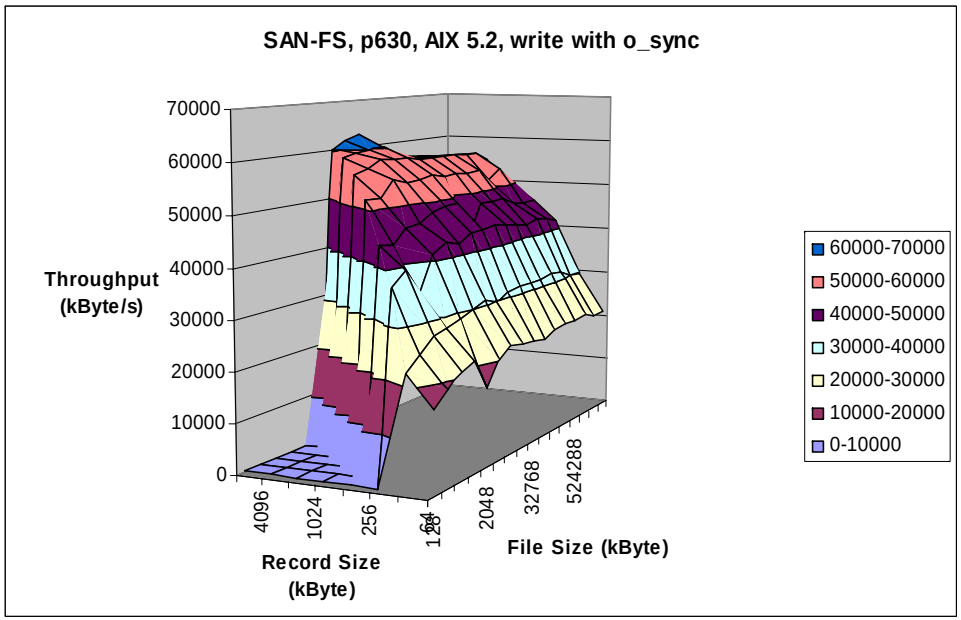




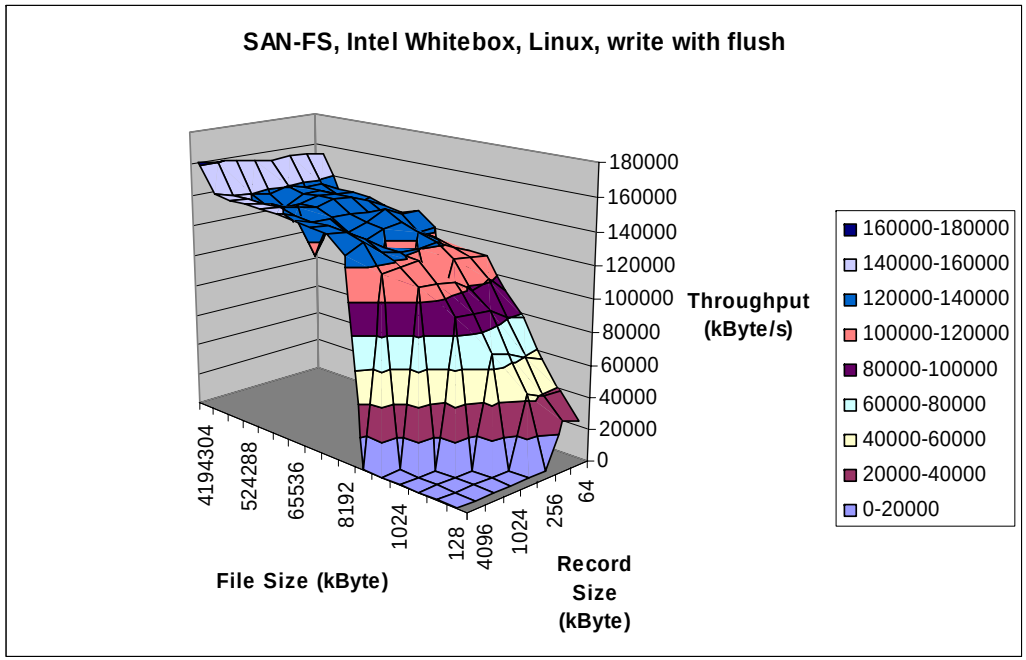
IBM SAN-FS Single Path

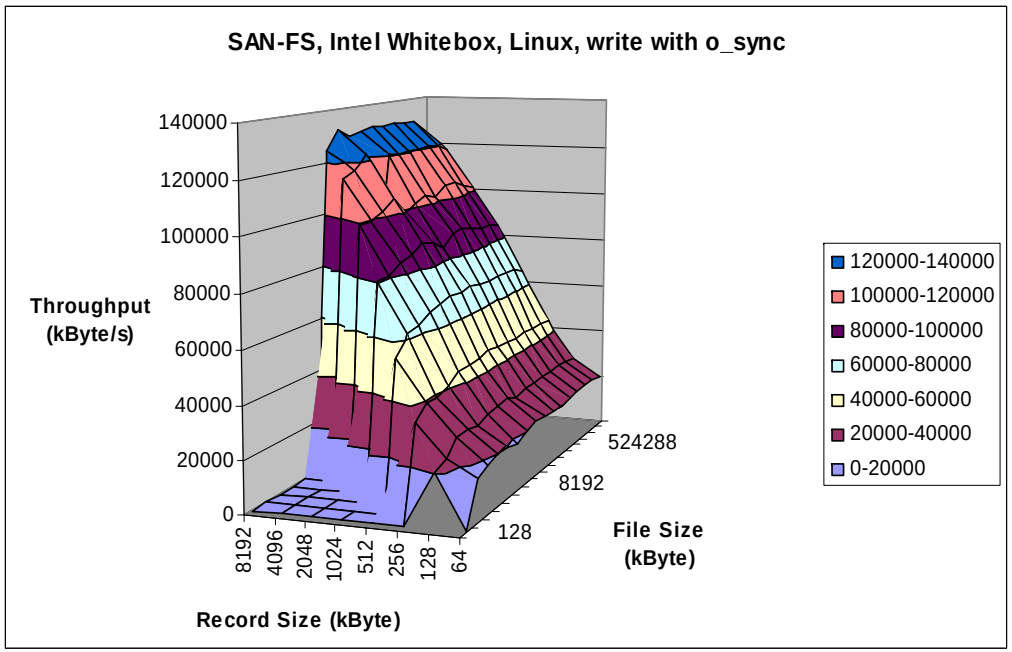




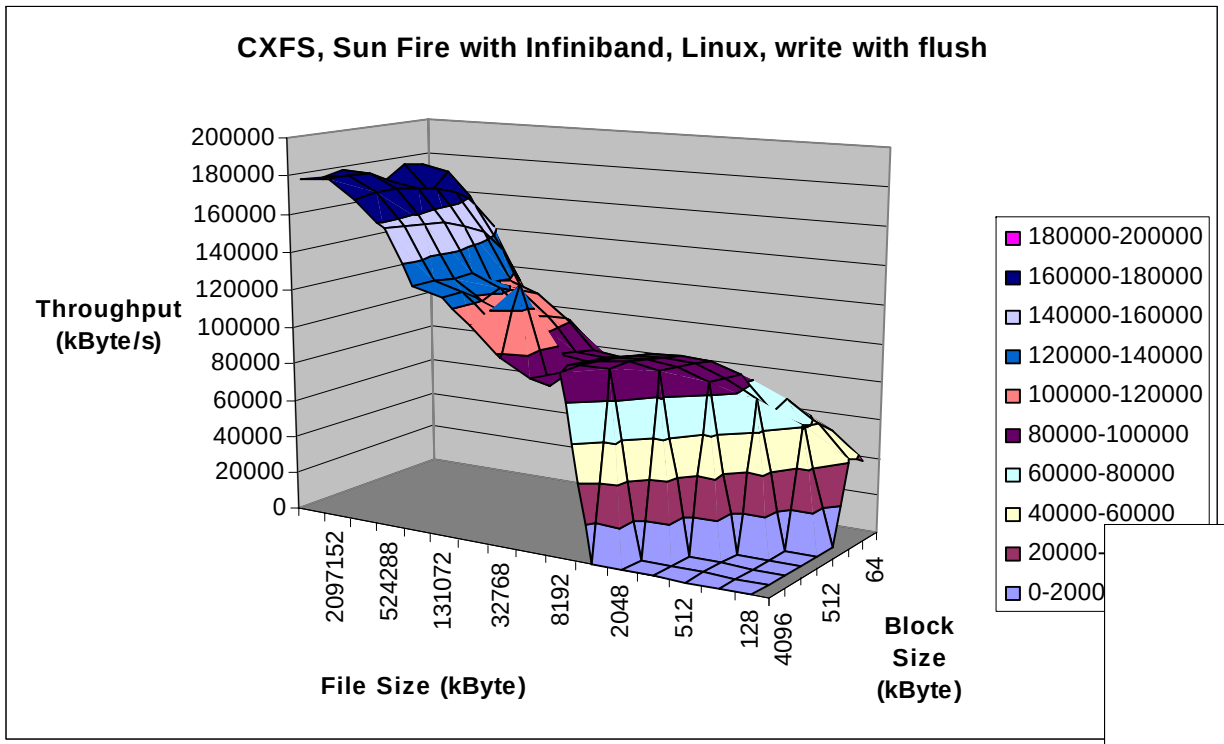


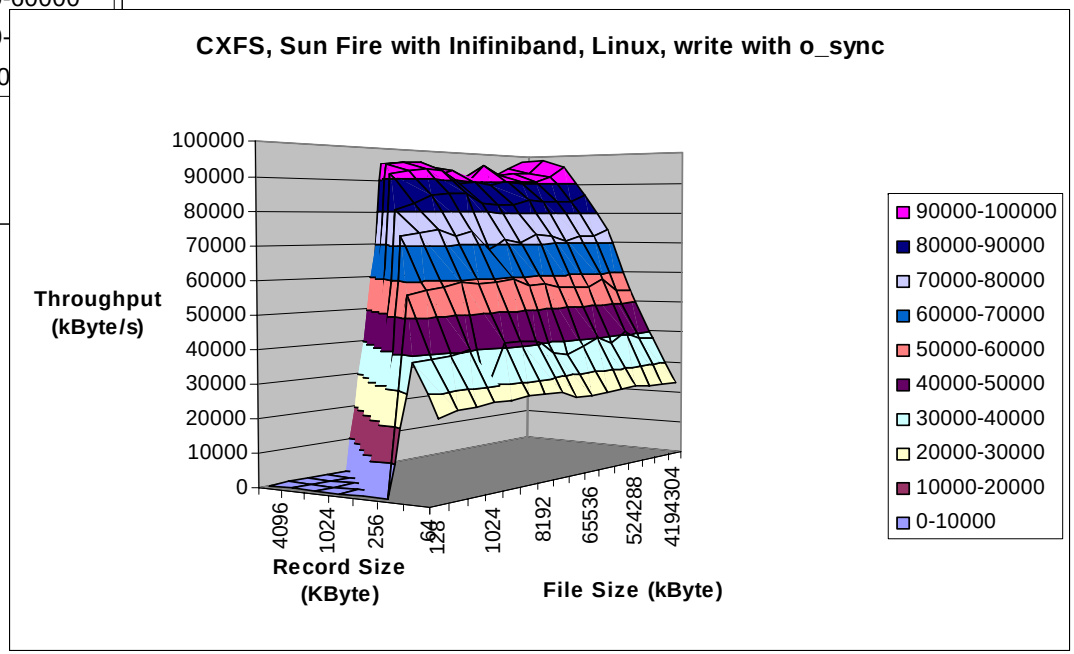
IBM SAN-FS Single Path



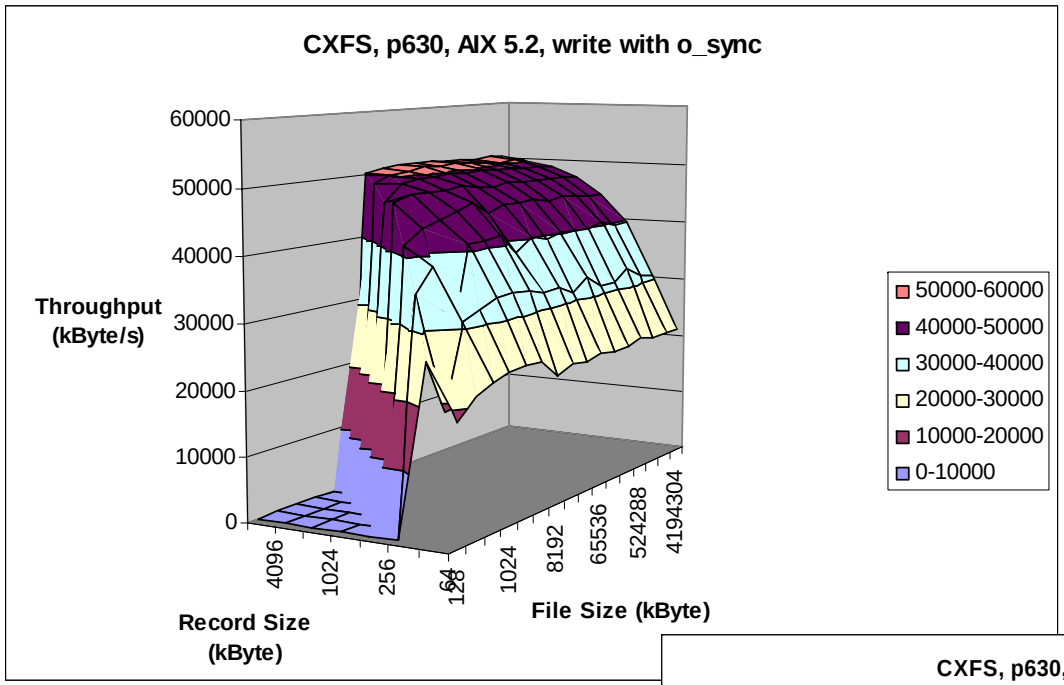


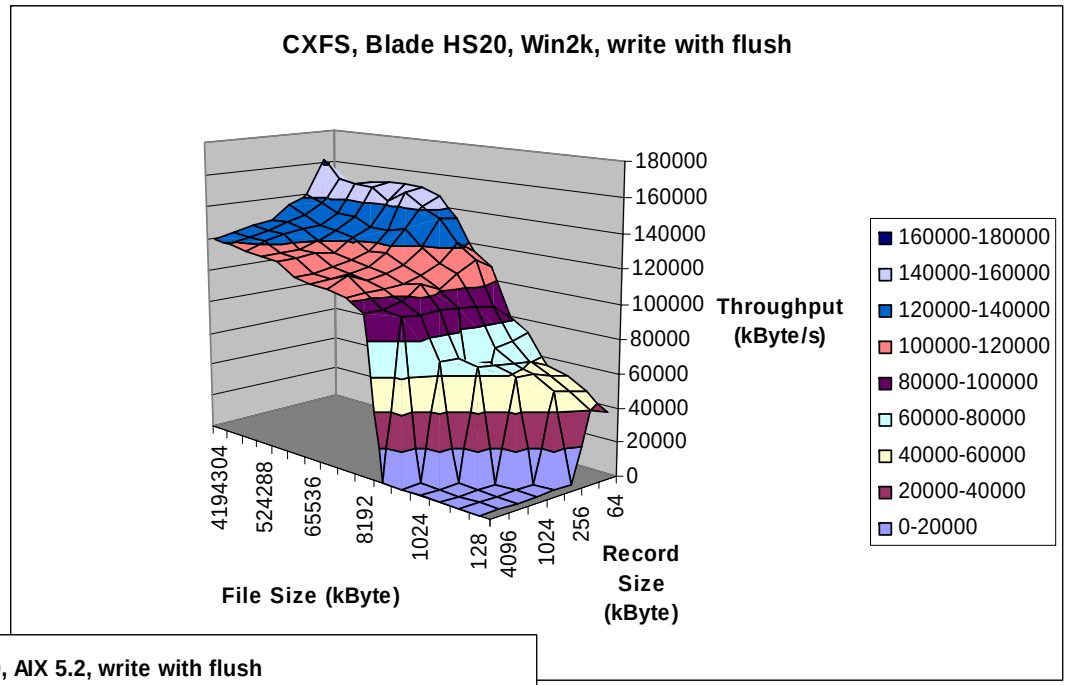
SGI CXFS Single Path

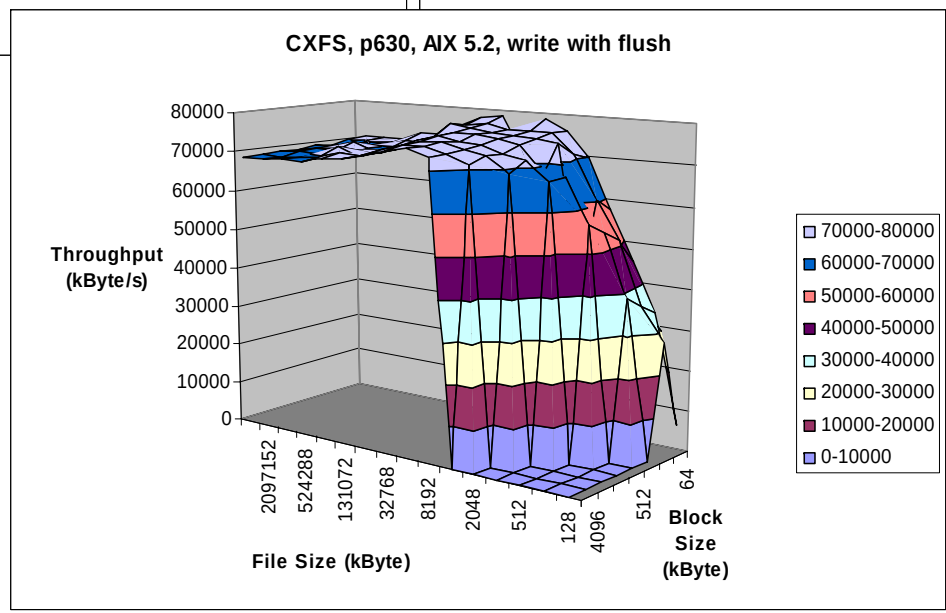




SGI CXFS Single Path

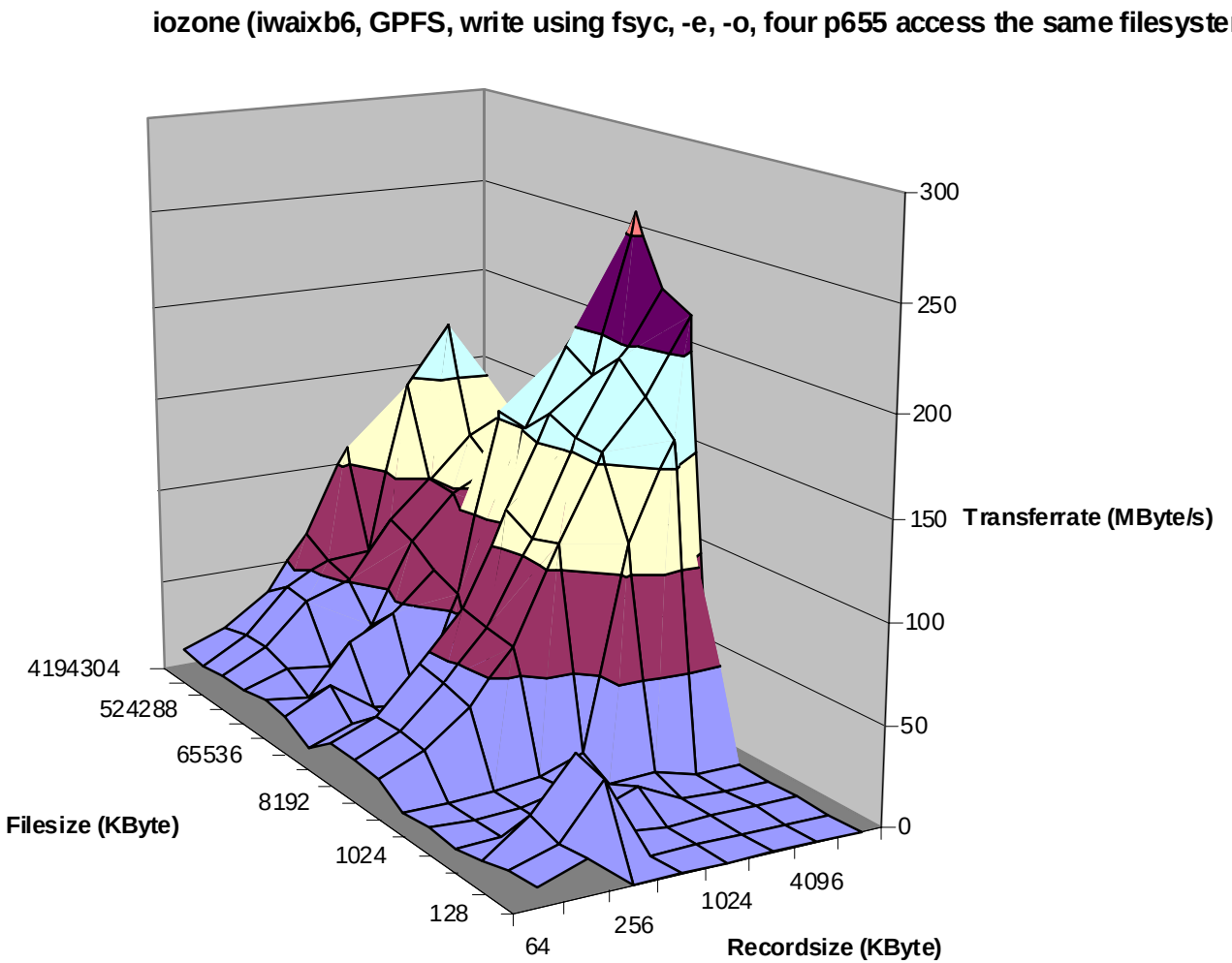






Results using a part of the production environment (GFPS)

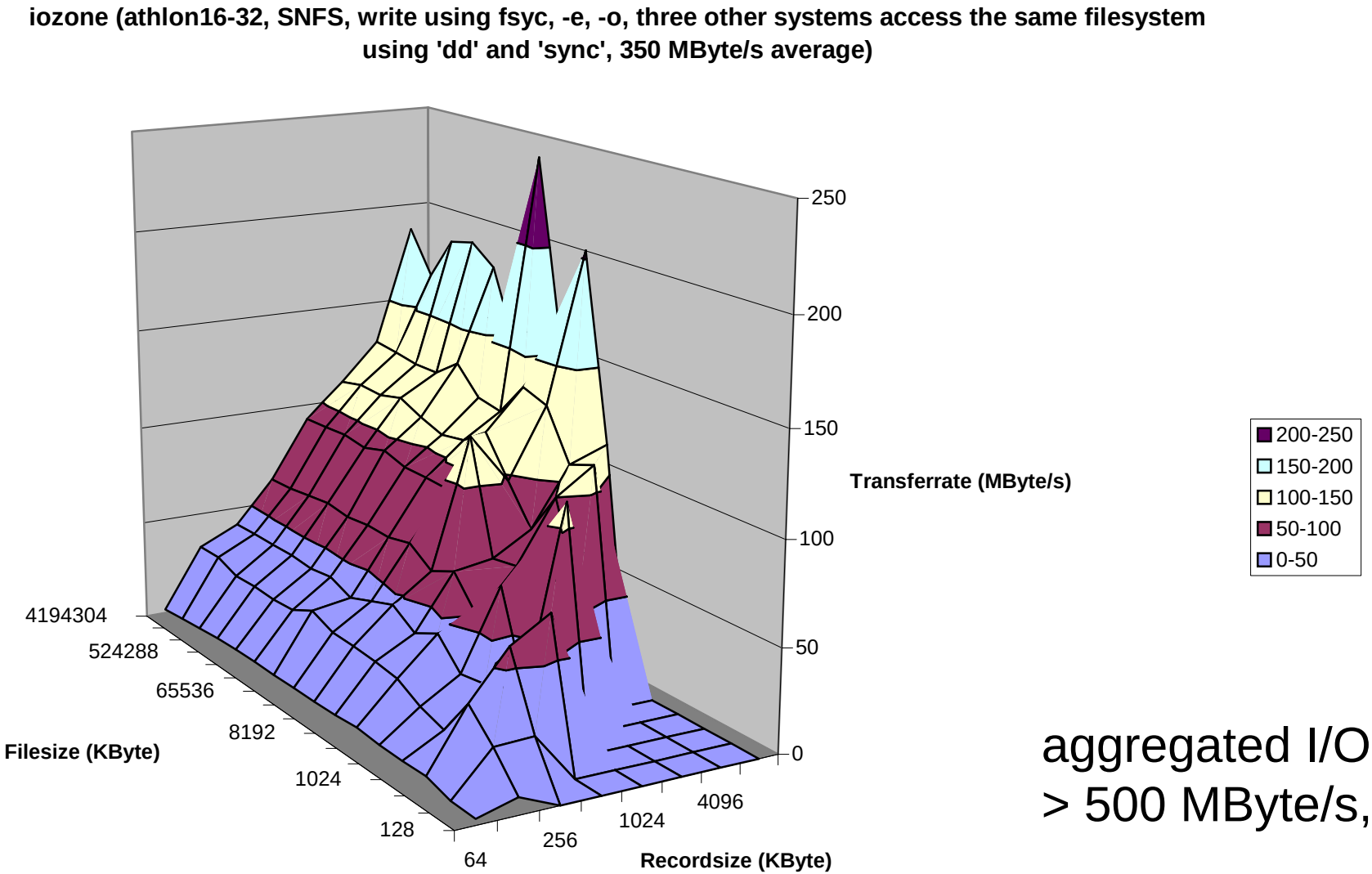
(4xp655, 2 HBAs each, 2 FastT700 systems, 2 connections each)



4 iozone tests running in parallel, aggregated transfer rate > 600 MByte/s

Results in the CampsGrid Testbed

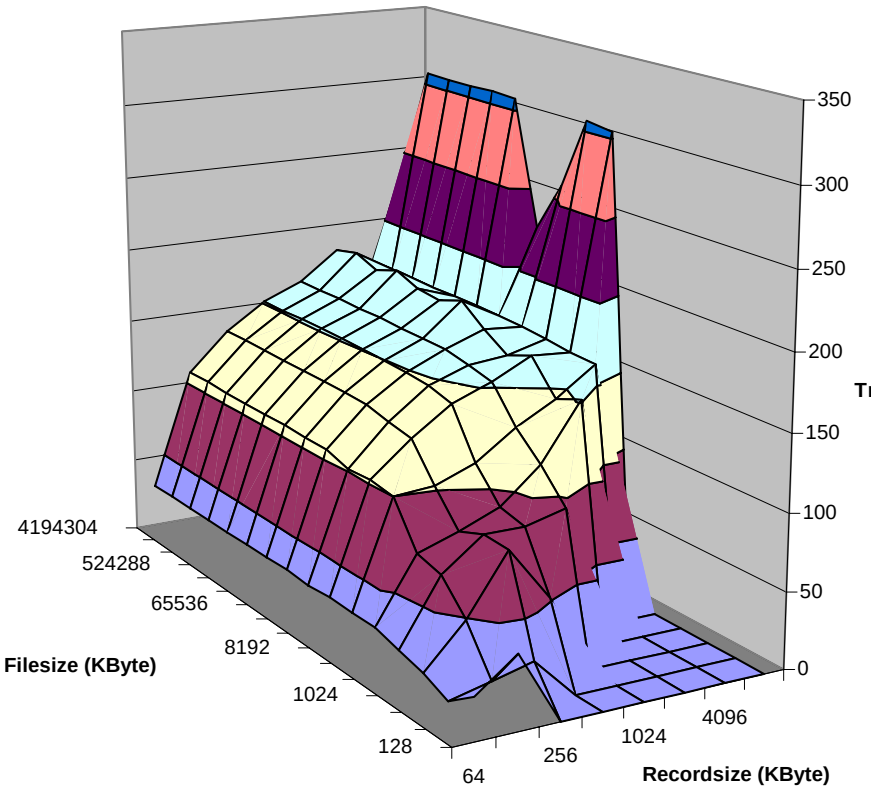
(SNFS, 350 MByte/s average traffic from three different systems)



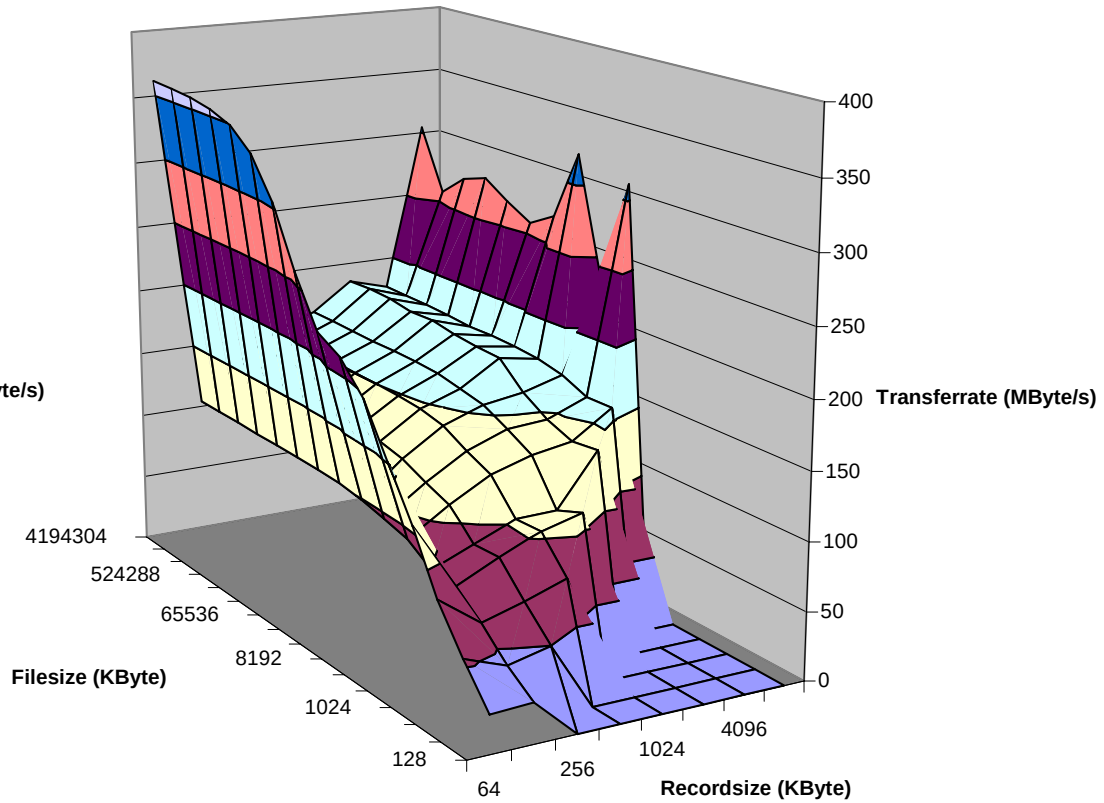
aggregated I/O rate
> 500 MByte/s,

iozone on different operating systems using the same hardware and SNFS

iozone (iwrcgblade1, SNFS, write using fsyc, -e, -o, Linux, 2 connections to DDN)



iozone (iwrcgblade3, SNFS, write using fsyc, -e, -o, Windows2000, two path test to DDN)



fsync seems to be not working for Windows2000 !!

High-Road Lösung von EMC² wird gerade getestet → Hard- und Software

SAN-FC-S-ATA Ausschreibung wurde vor wenigen Tagen von FSC/EMC² gewonnen (CX700)→ Preis/Leistungsverhältnis bei DDN war für CampusGrid nicht akzeptabel

SVC-Lösung von IBM zur Virtualisierung nicht wirklich gut!

Momentan hat SNFS wegen der Einfachheit und der Leistungsfähigkeit im Projekt CampusGrid die Nase vorn.

20 Clienten pro MDS sind bei SNFS unter Volllast kein Problem