Stephan Wiesand
DESY - DV-
2006-12-19

The Panasas Storage System

at

Standort Zeuthen

Outline

- Why another storage solution?
- Technical Description
- Performance Measurements
- Availability & Usage

Current Storage Mix

- AFS
  - general purpose, accessible from any system, secure
  - scales very well - if:
    - datasets distributed across volumes
    - volumes distributed across file servers
    - access patterns match distribution pattern
    - (too?) much overhead for transient datasets
    - global namespace, distribution by volume (manual)
- dCache
  - fast & scalable, but not general purpose
    - large, static files only (files can not be modified)
    - requires preload library or special API to access
    - global namespace, distribution by file (automatic)
- NFS
  - where users are unable to use anything else, or simply insist

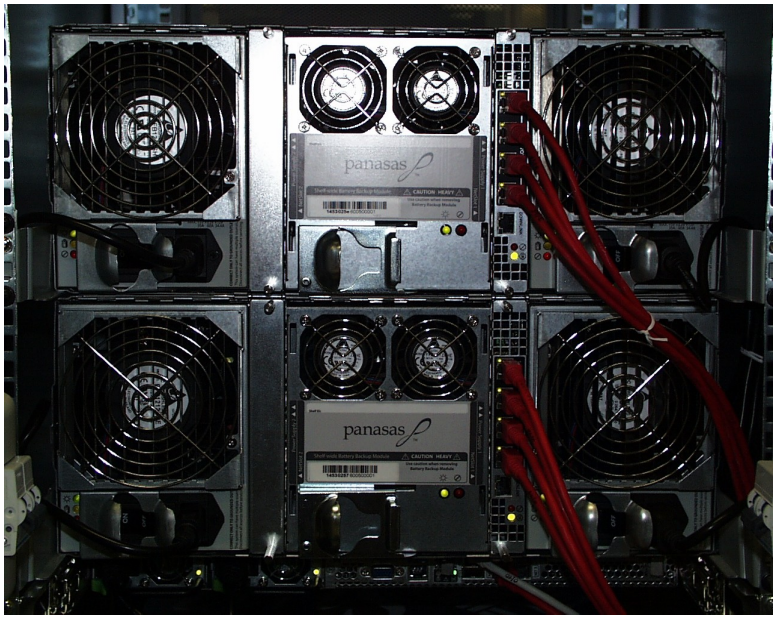
What's Missing

- some amount of storage that
  - can be used from many clients in parallel
    - dozens to hundreds
  - performs well
    - several hundred MB/s
  - behaves like an ordinary file system
    - without a need for special access methods
  - looks like a single blob of space
    - without a need to distribute data manually, or even think about it
  - is suitable for typical datasets (mixture of file sizes)
    - keeping millions and millions of very small files is abuse of any storage
- all at the same time

Panasas

- 11 blades per 4U shelf
- each blade is a complete system
- two flavors:
  - storage blades
  - director blades

- 1 Gigabit Ethernet Switch per shelf
- each has 4 aggregated GbE uplinks
- redundant power supplies and fans



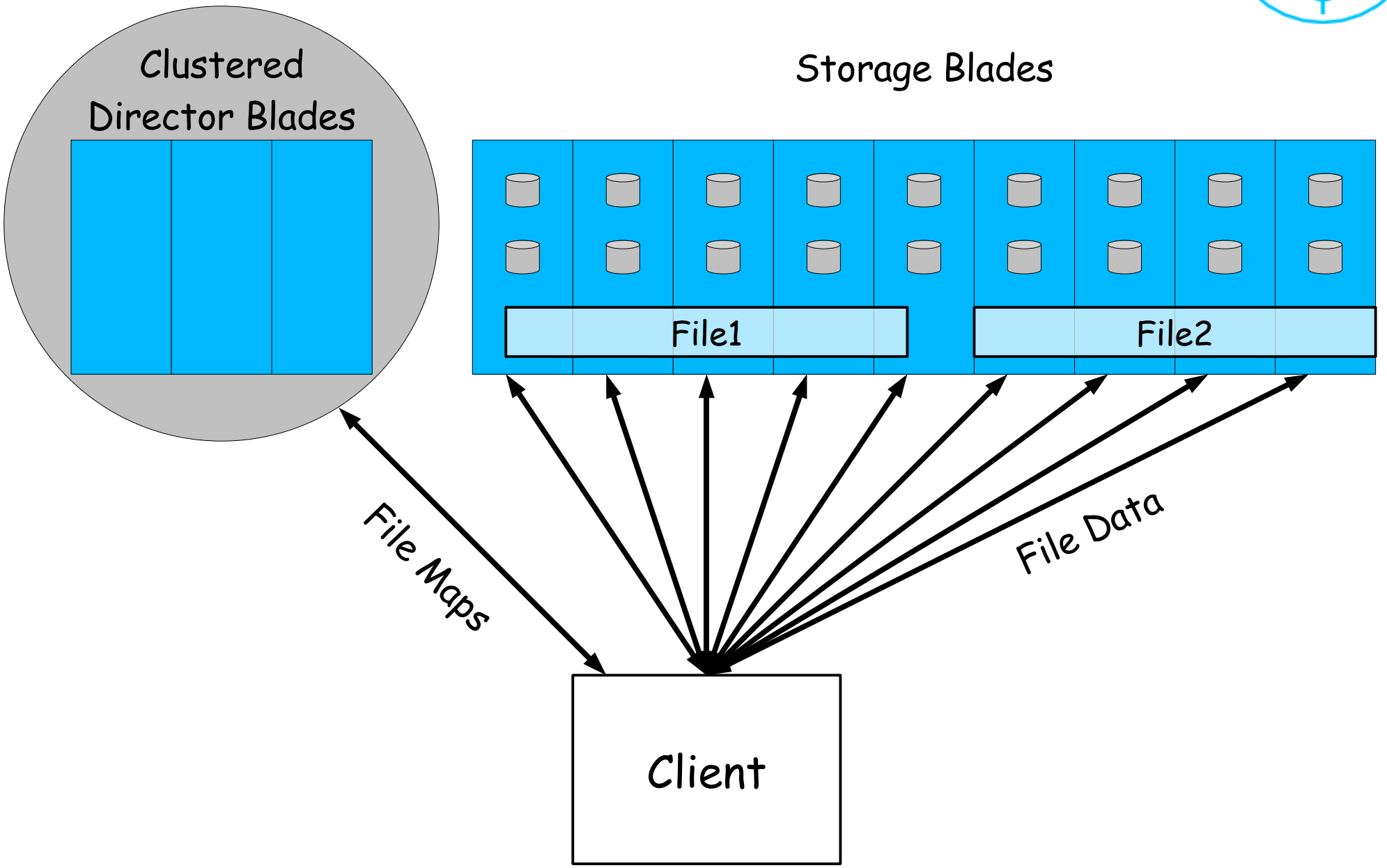
Details

- Storage Blades:
  - 2 SATA data disks,Celeron CPU, 512MB RAM
- Director Blades:
  - single system disk, Xeon, 4 GB RAM
- ActiveScale Operating System
  - FreeBSD + Storage Cluster Softwares
- Data is distributed across storage blade disks automatically
  - small files: mirrored
  - larger files: N+1 stripes for data + parity (RAID-5 like)
  - director blades keep a map for each file

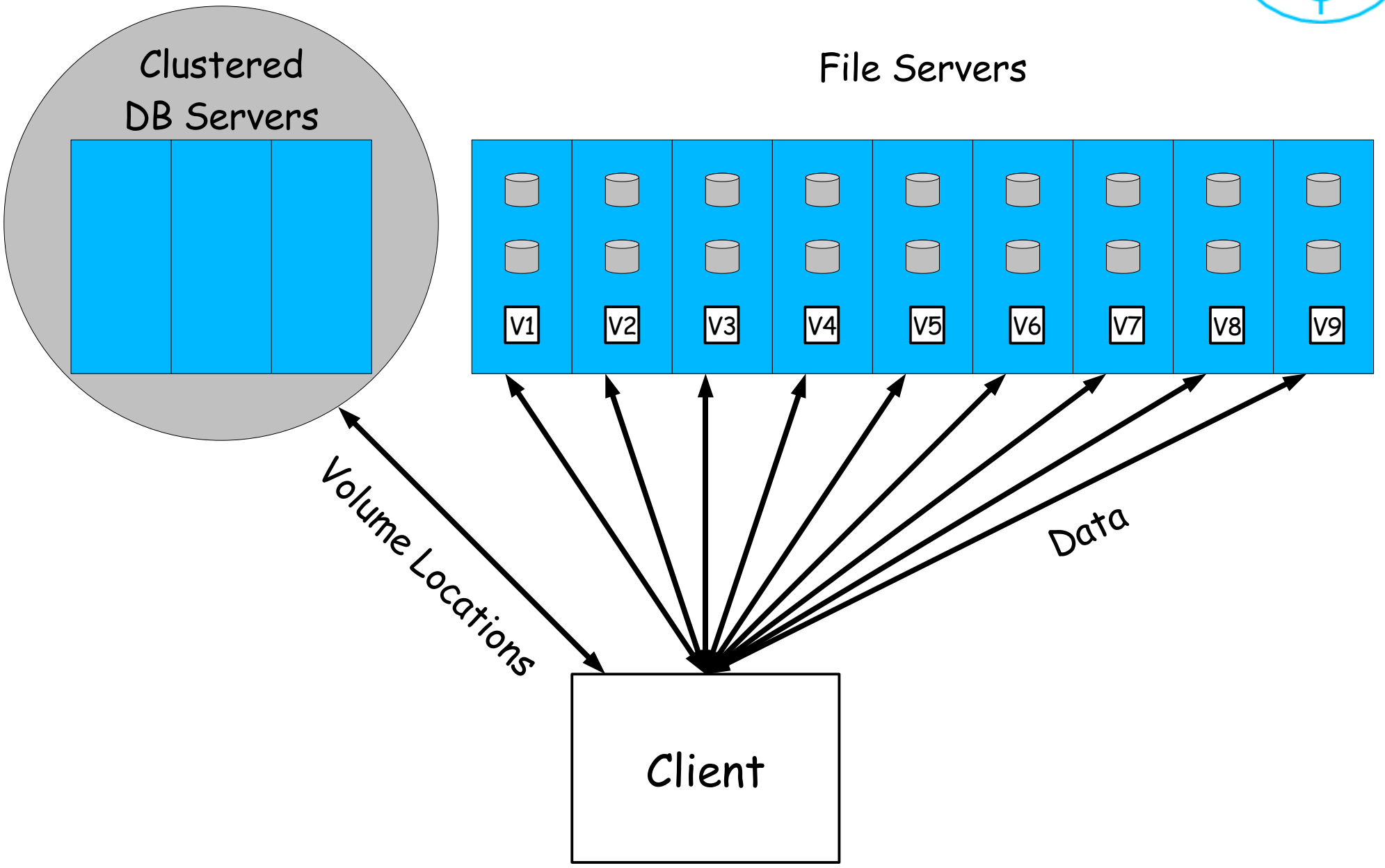
Client Access

- either through director blades
  - CIFS (Samba)
  - NFS (V3)
- or through DirectFlow client
  - obtains file distribution map from director blades
  - reads/writes data directly from/to right storage blade
  - available for major linux distributions
    - ports to custom kernels possible
- security: like NFS
  - must trust client system
  - no Kerberos tickets/tokens

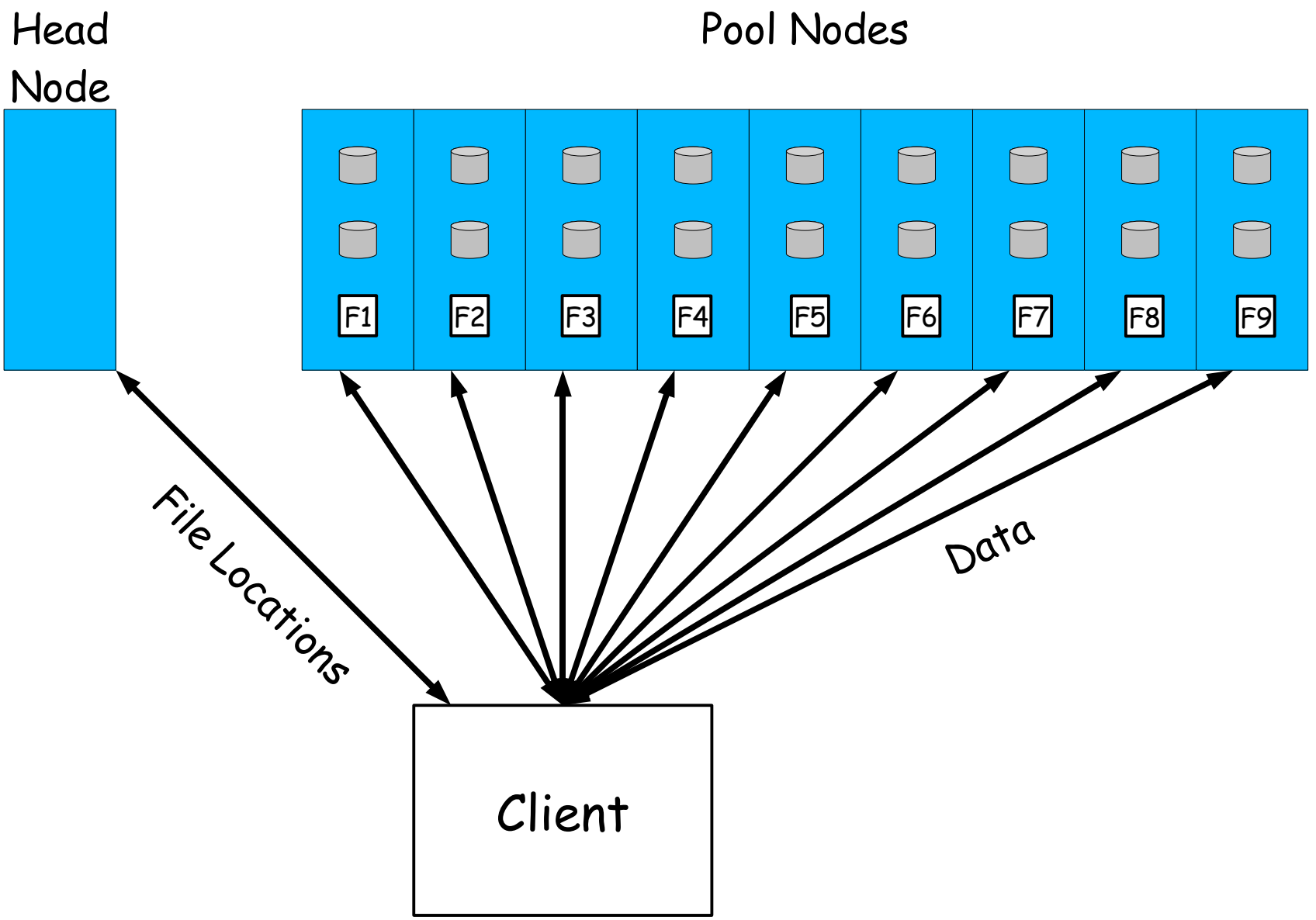
DirectFlow Client Access



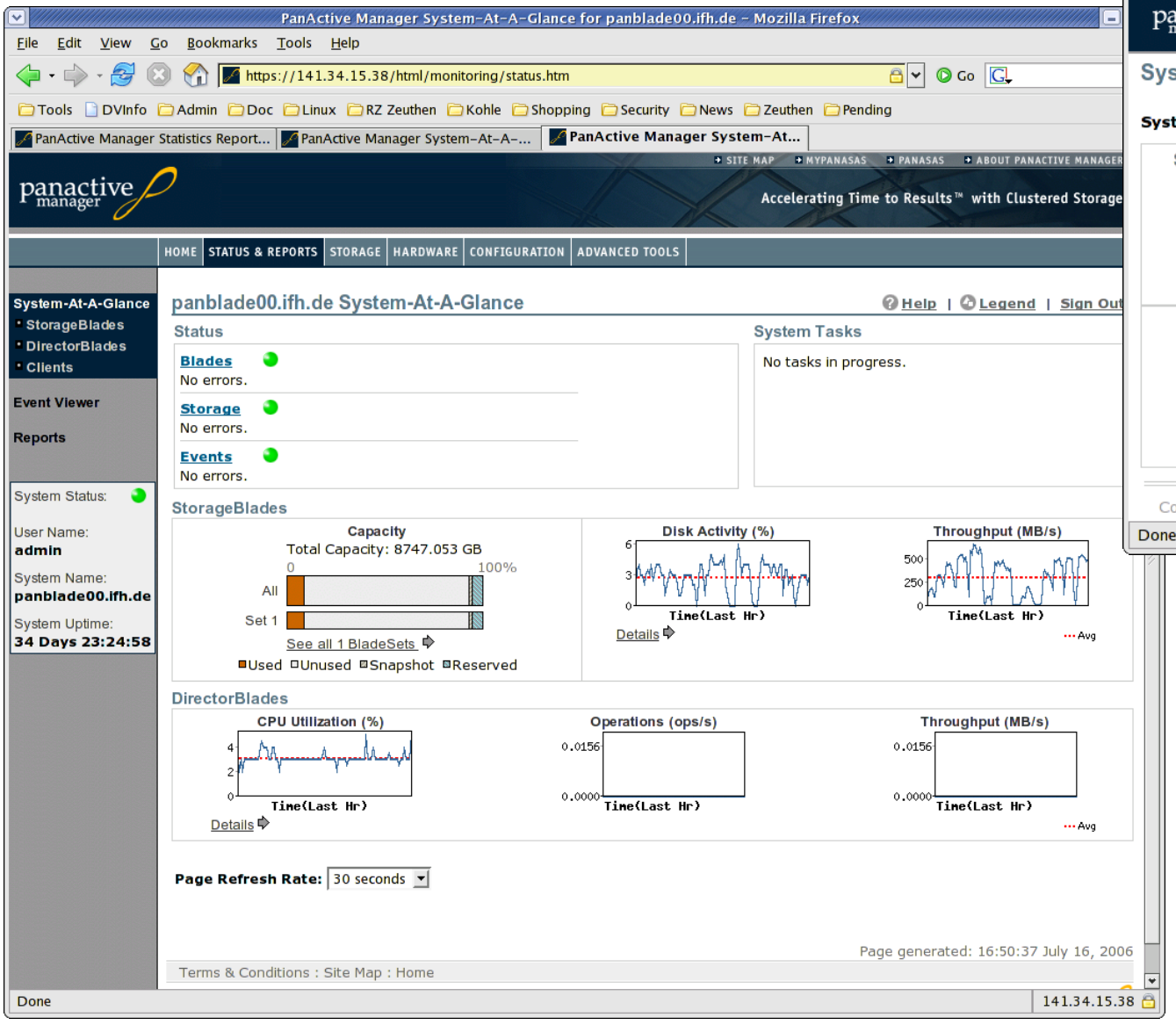
NB: AFS

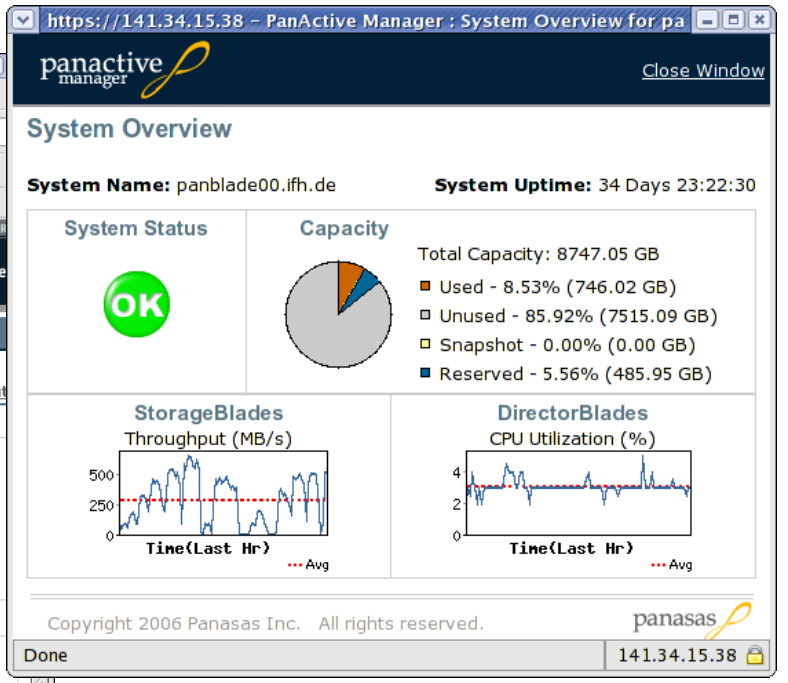

Clustered
DB Servers
File Servers
V1
V2
V3
V4
V5
V6
V7
V8
V9
Volume Locations
Data
Client

NB: dCache

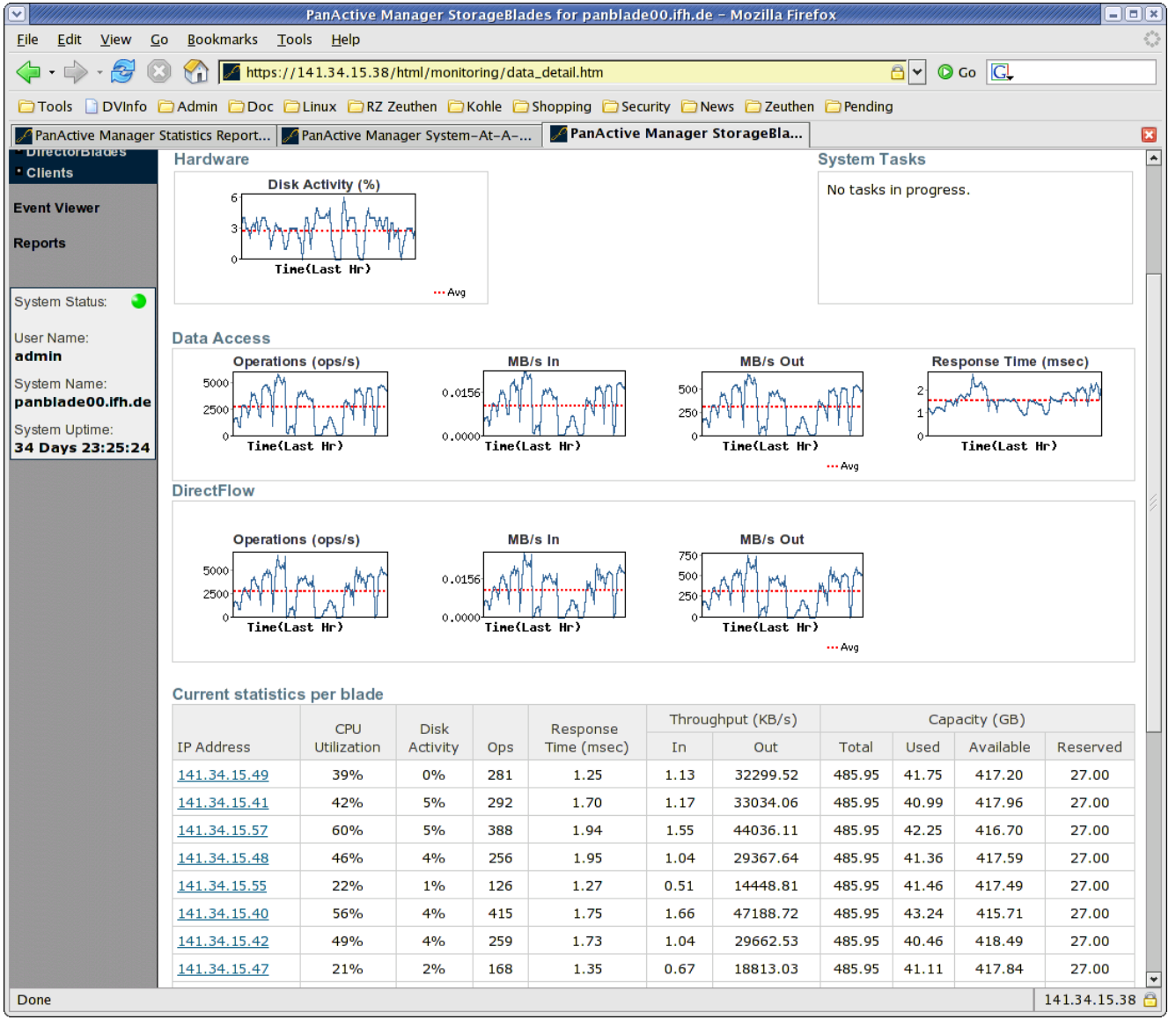

Head Node
Pool Nodes
F1
F2
F3
F4
F5
F6
F7
F8
F9
File Locations
Data
Client

Web Interface



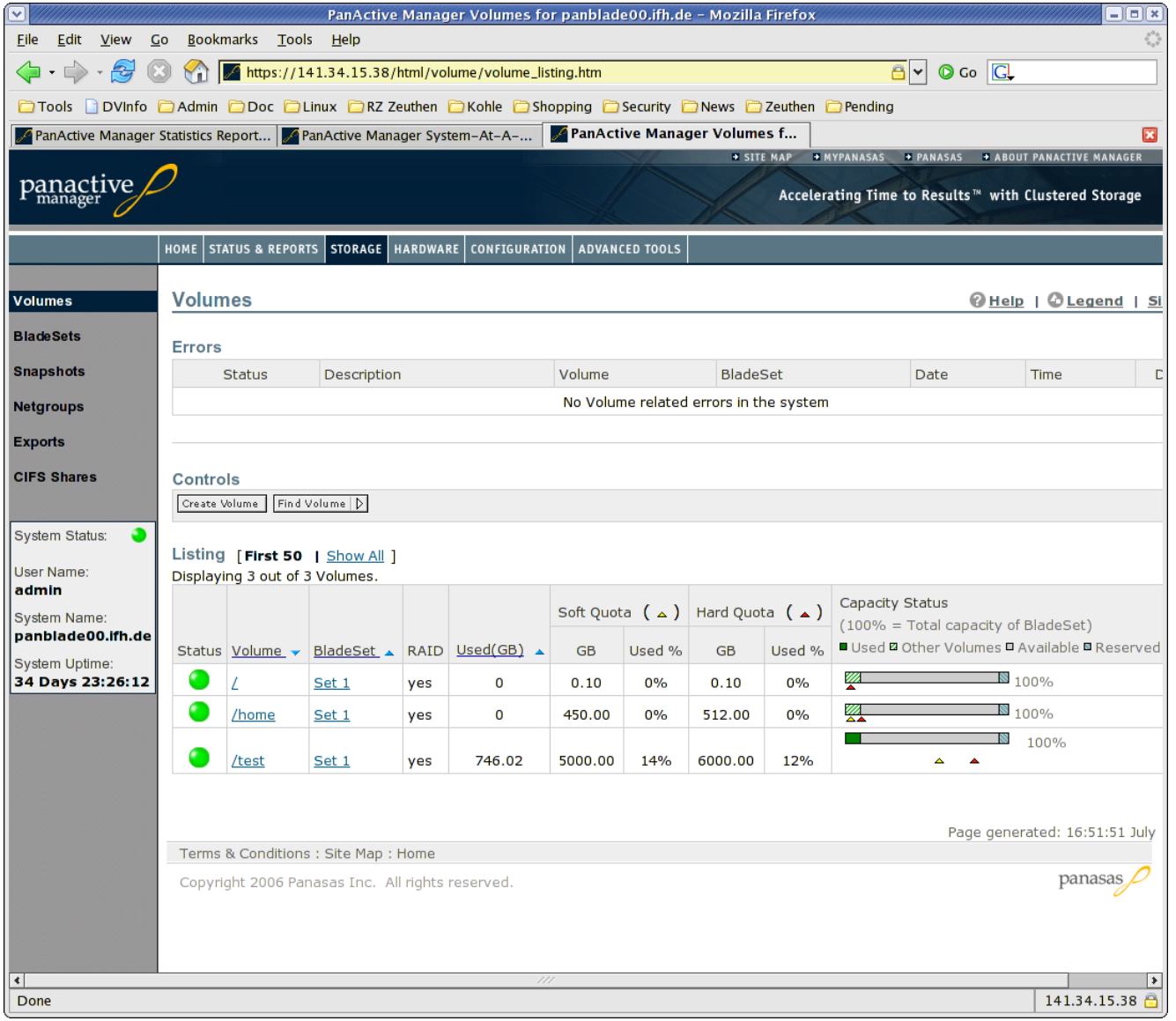


Web Interface: Performance

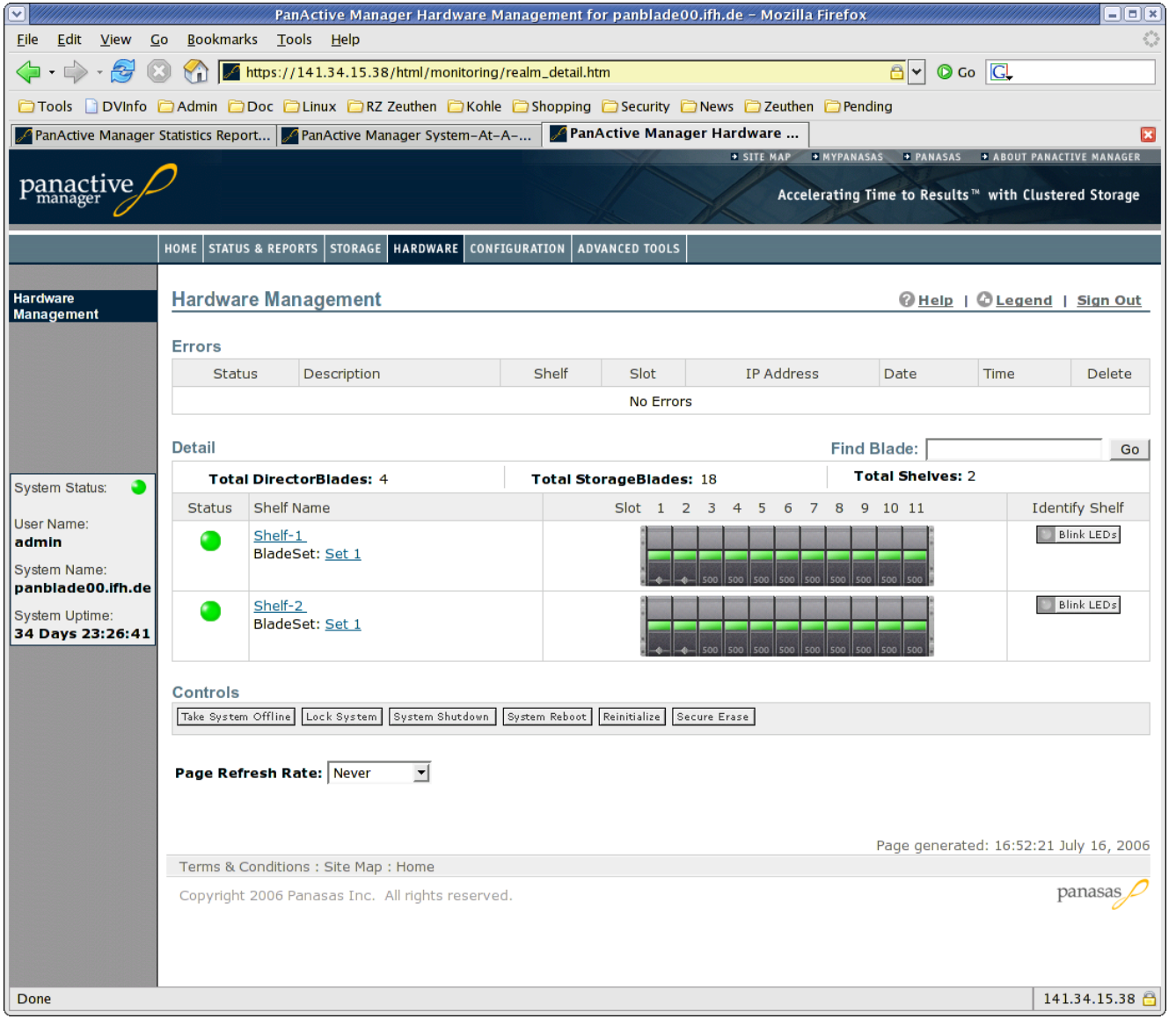


IP Address	CPU Utilization	Disk Activity	Ops	Response Time (msec)	Throughput (KB/s)		Capacity (GB)			
					In	Out	Total	Used	Available	Reserved
141.34.15.49	39%	0%	281	1.25	1.13	32299.52	485.95	41.75	417.20	27.00
141.34.15.41	42%	5%	292	1.70	1.17	33034.06	485.95	40.99	417.96	27.00
141.34.15.57	60%	5%	388	1.94	1.55	44036.11	485.95	42.25	416.70	27.00
141.34.15.48	46%	4%	256	1.95	1.04	29367.64	485.95	41.36	417.59	27.00
141.34.15.55	22%	1%	126	1.27	0.51	14448.81	485.95	41.46	417.49	27.00
141.34.15.40	56%	4%	415	1.75	1.66	47188.72	485.95	43.24	415.71	27.00
141.34.15.42	49%	4%	259	1.73	1.04	29662.53	485.95	40.46	418.49	27.00
141.34.15.47	21%	2%	168	1.35	0.67	18813.03	485.95	41.11	417.84	27.00

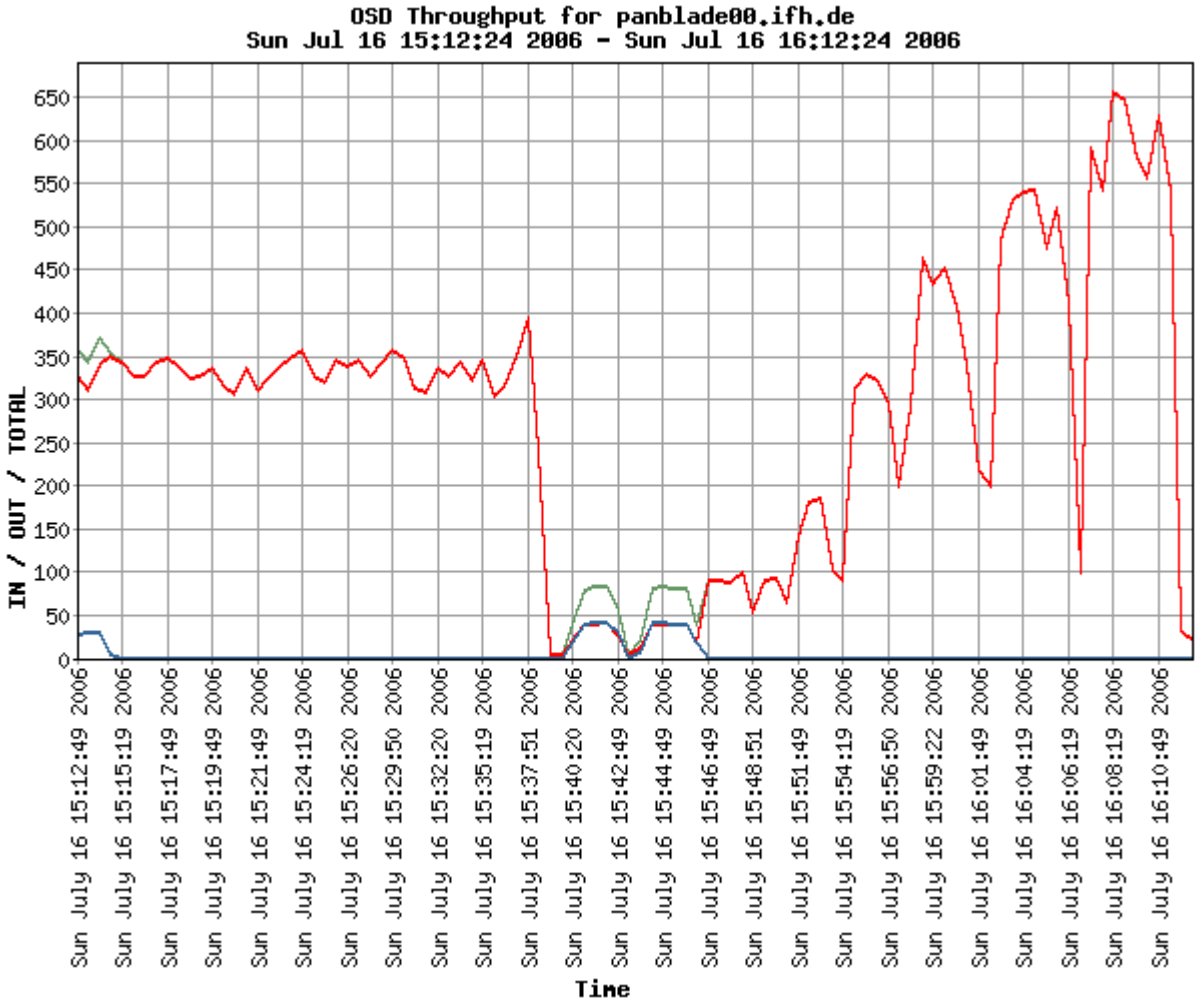
Web Interface: Volume Management



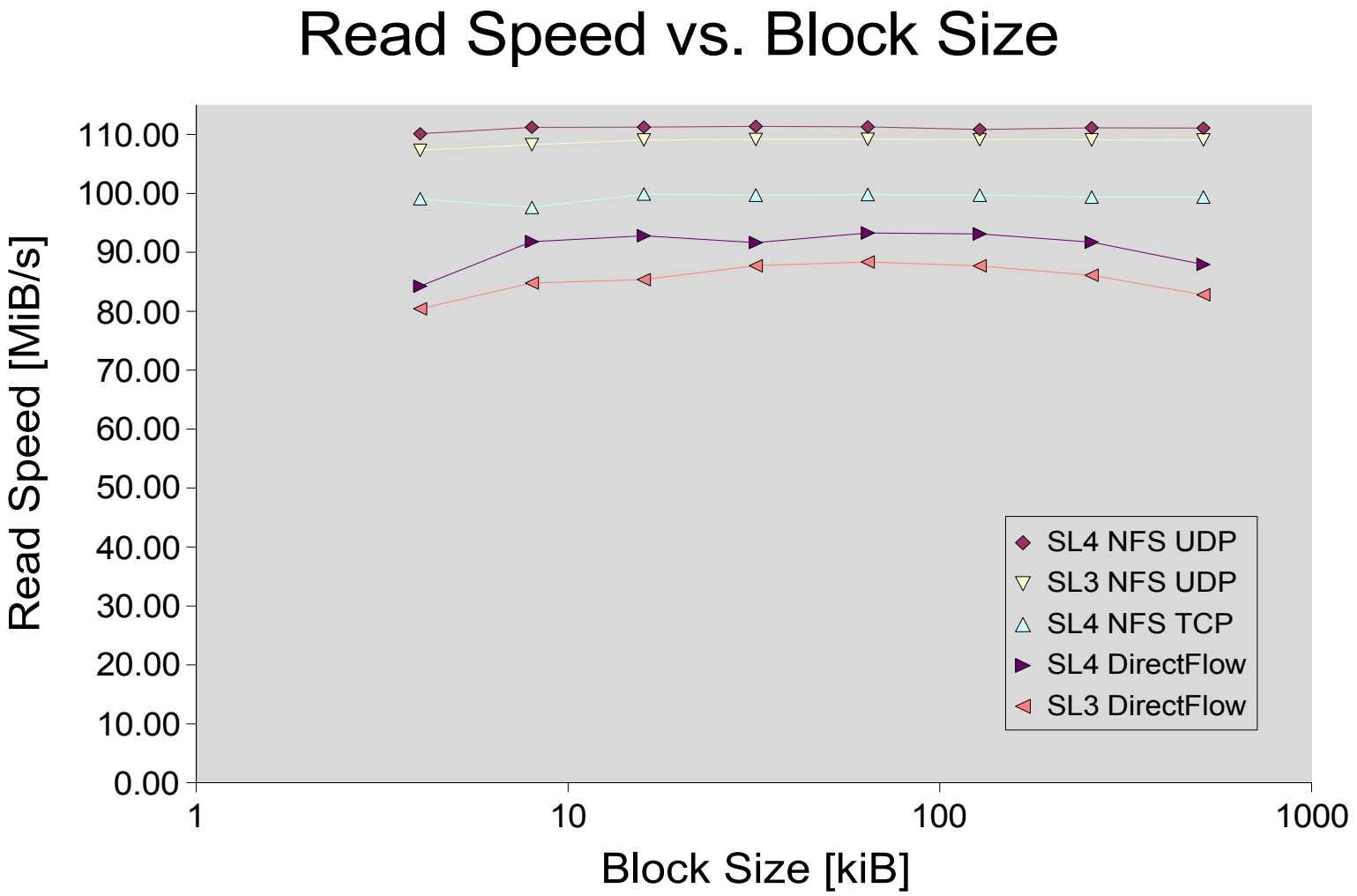
Web Interface: Hardware



Web Interface: Throughput

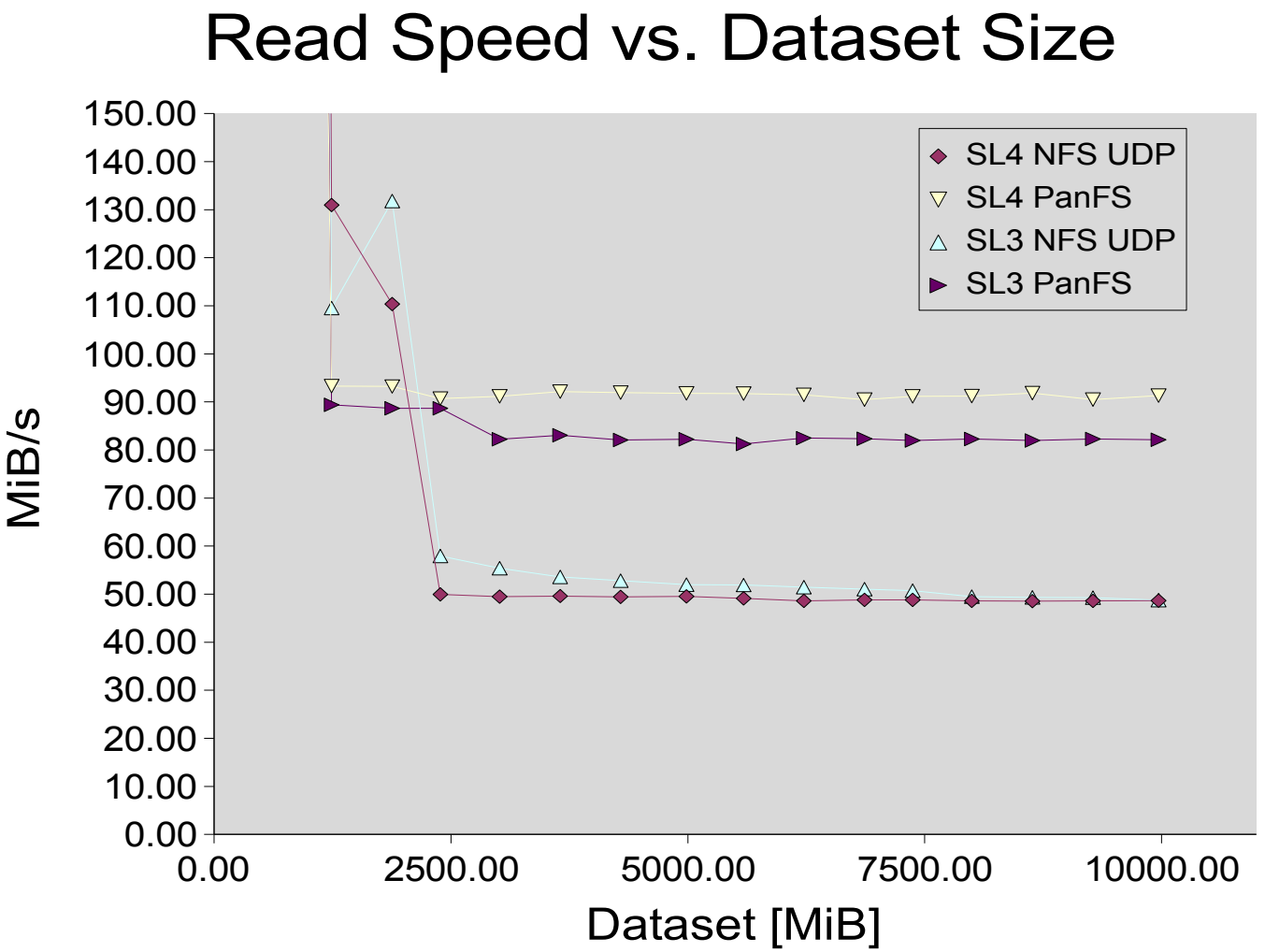


Performance: Single Client

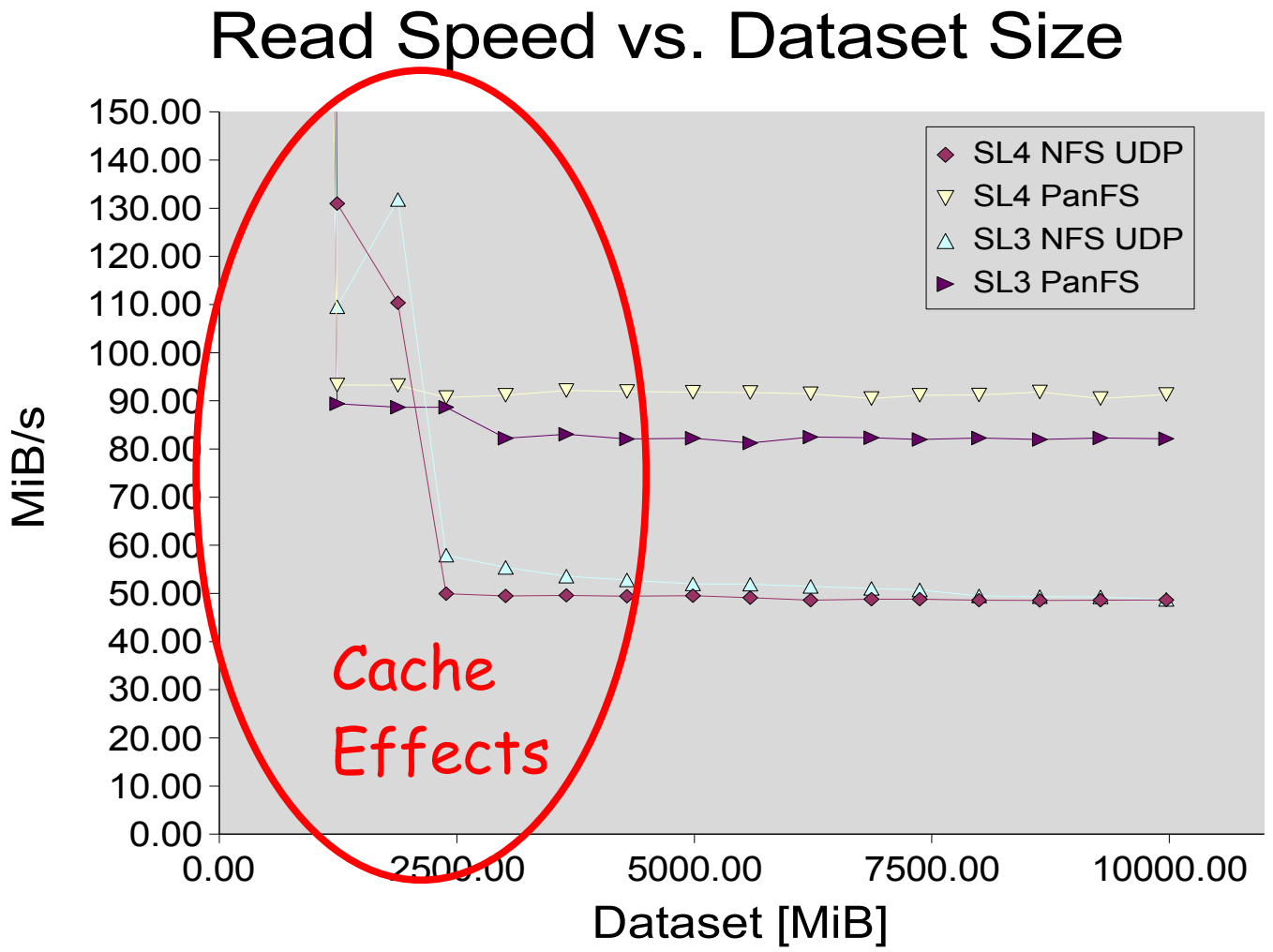


- following synthetic tests: DirectFlow client, 64 kiB request size

Performance: Single Client



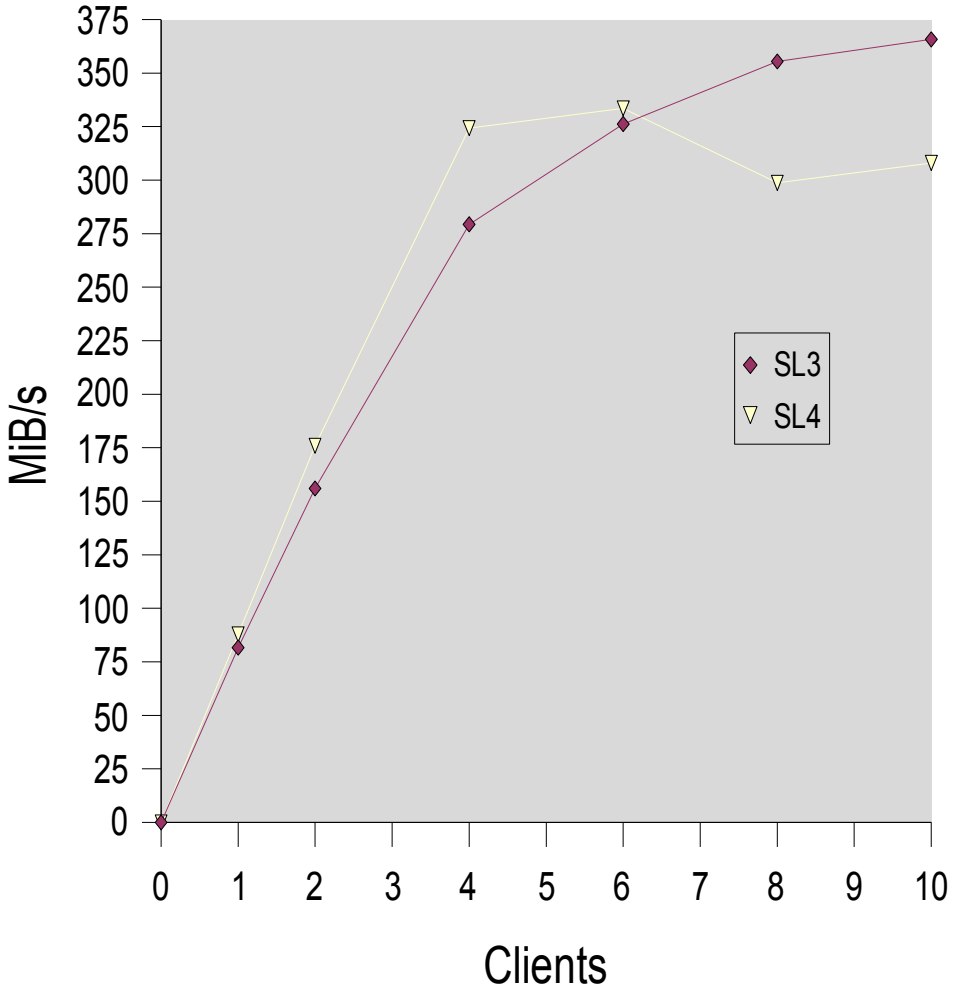
Performance: Single Client



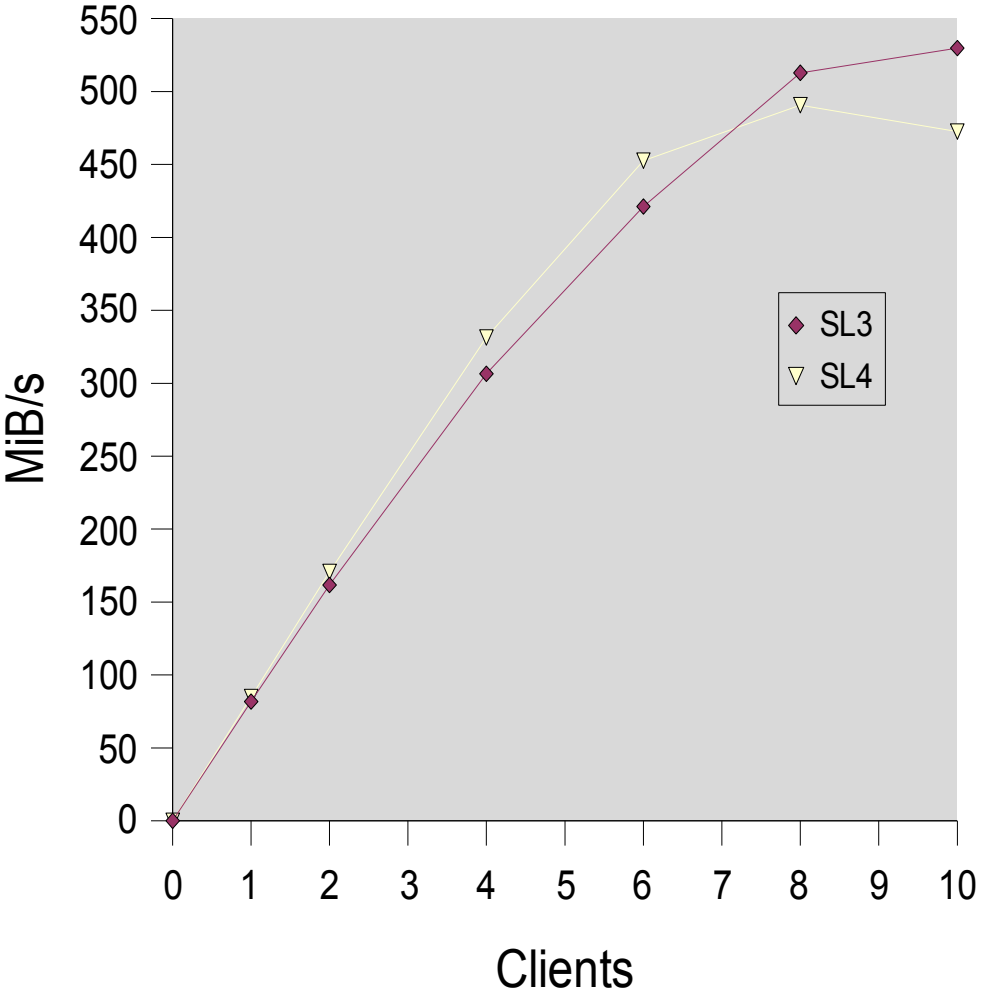
- Dataset for synthetic tests: ~ 5 GB

Performance: Multiple Clients

Read Throughput, 1 Shelf

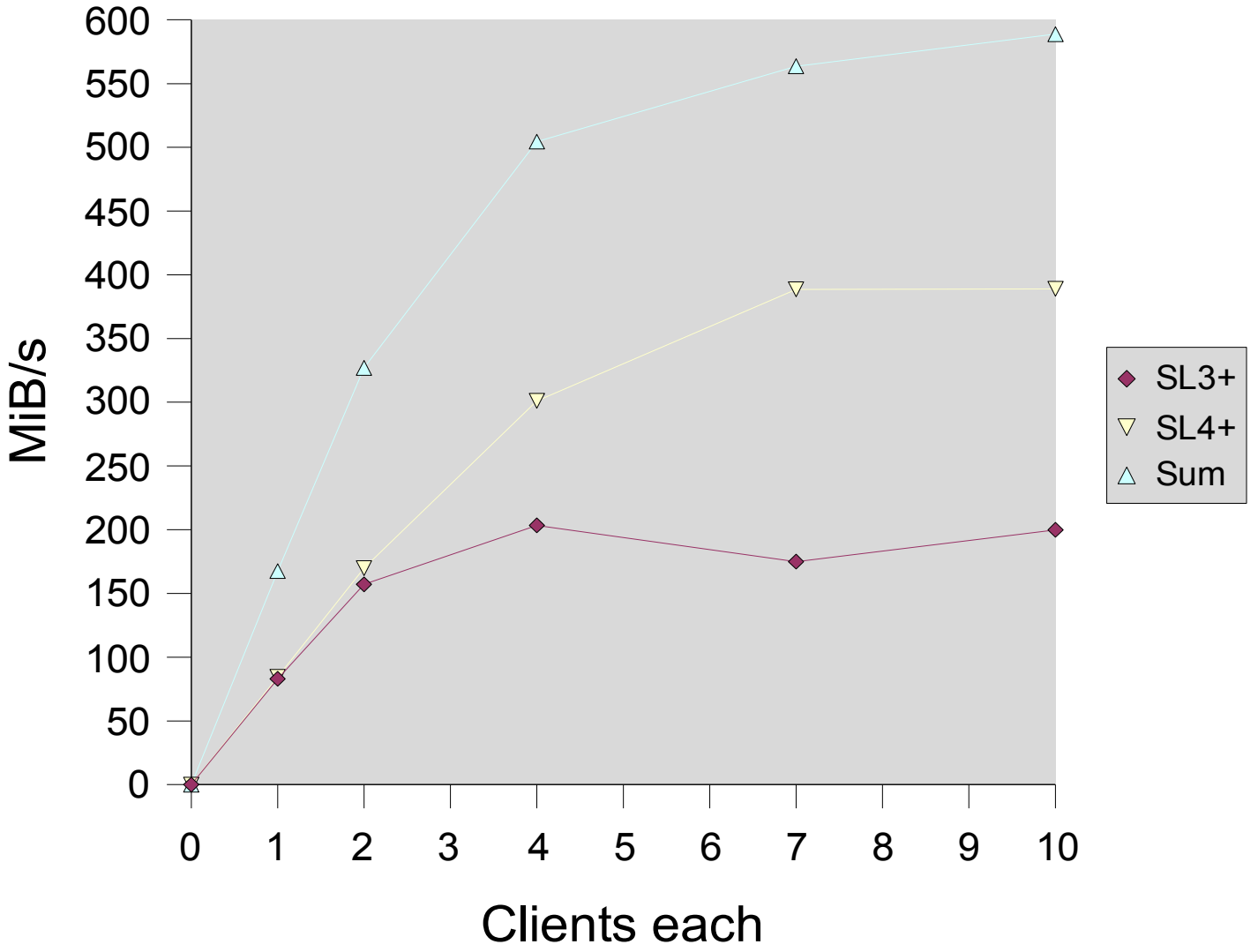


Read Throughput, 2 Shelves

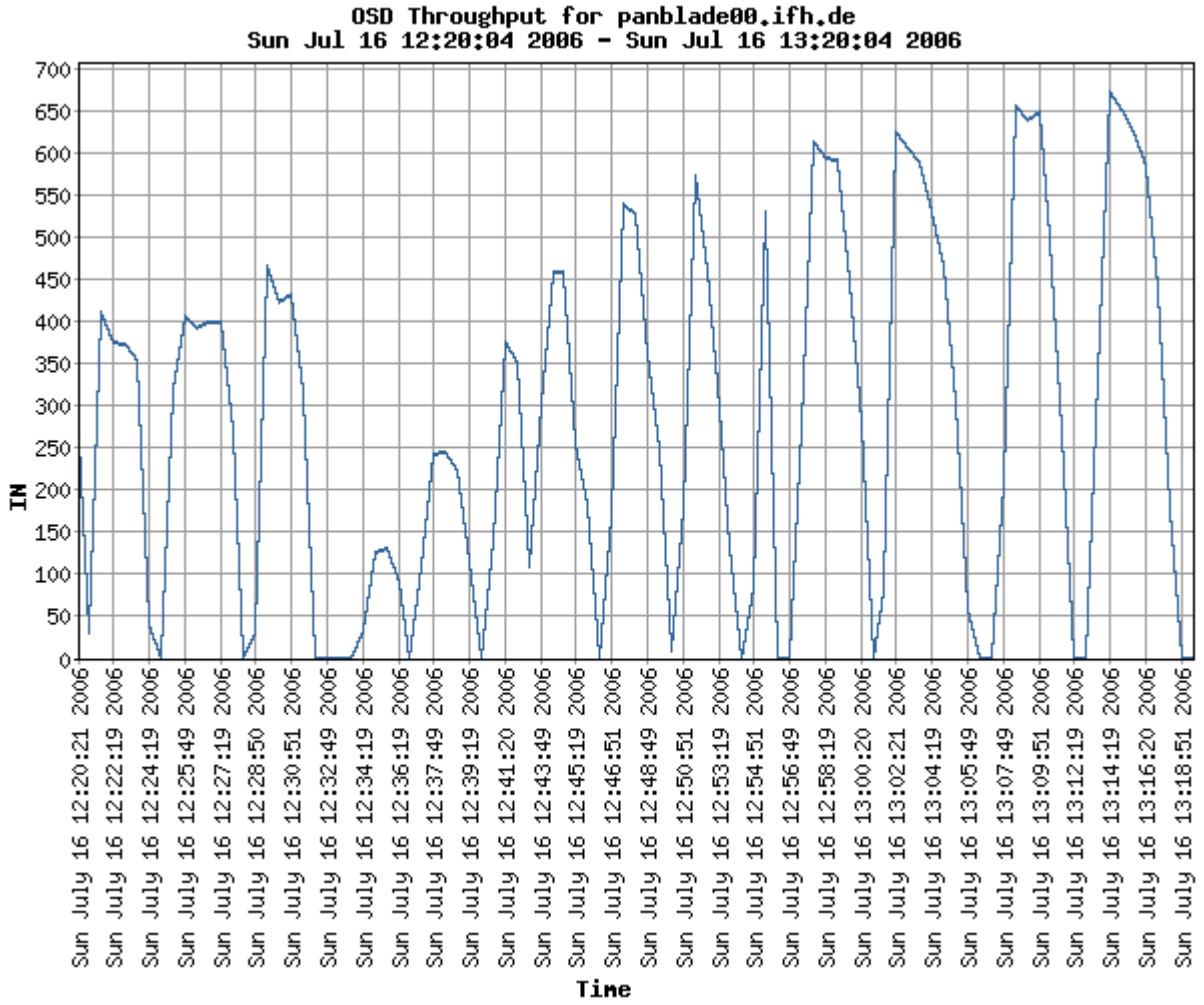


Performance: Multiple Clients

Read Throughput, 2 Shelves, SL3/4 Mix



Write Throughput, up to 20 Clients



Performance Tests: Summary

- panasas test system had 18 storage blades
- test clients were galaxy11-30
  - connected to same switch as the panasas system
- up to ~ 600 MiB/s payload in synthetic tests
  - read & write
  - between client & disk platter
    - care was taken to avoid measuring the cache
- up to ~280 MiB/s observed in real life use
  - systems not all connected to same switch
  - network may have been bottleneck

Near Term Future

- test system was purchased very recently
- final system will have 19 storage blades + 3 director blades
  - -> clustering the directors, access through DirectFlow
- net capacity: ~8 TB
- system will soon be
  - connected to a dedicated, non-routed subnet
    - no bandwidth load on regular network
    - allows use for Tier2 VO-space
  - updated to latest ActiveScale release
    - 3.0, still beta, but close to final
    - should be final when system run-in with final setup

Usage & Availability

- available on farm, transfer, WGS-like systems
- structure foreseen:
  - /panfs/group/<group>/<project>
  - /panfs/group/<group>/user/<user>
  - volumes will have to be created/deleted by admins (uco)
    - cli available, hence an afs_admin like solution is possible, but would need to be implemented
- usage: nothing special:
  - except: du -> pan_du, df -> pan_df
  - quotas (soft/hard) per volume, e-mail alerts when exceeded
  - no ACLs; no token required

Summary

- the panasas system adds to the storage mix a limited amount of space that's
  - easy to use
  - very performant when accessed by many clients
- volumes on test system available on request
- client could be installed on additional systems
  - running an SMP kernel
  - physically located in a trusted area
  - centrally maintained, w/o root access for users/group admin
- final setup soon
  - data from test setup can probably not be kept

Stephan Wiesand
DESY - DV-
2006-12-19

Scientific Linux 4 & 5

at

Standort Zeuthen

Outline

- SL4: available now
  - what's new
  - what's not
- SL5: available soon
  - status
  - anticipated schedule
- migration SL3 -> SL4/5

But SL3?

- production system since January 2005, has been very stable
- still works on latest hardware
  - Dell 9G servers, SUN galaxy, latest Dell Precision Workstations
    - sound remains a challenge
- current release in Zeuthen: 3.0.7
  - 3.0.8 last minor update, will be rolled out in Zeuthen as well
- SL3 supported by FNAL/CERN until 10/07
- afterwards, if systems remain to be supported:
  - updates available from CentOS project
  - or RHEL3 subscriptions could be purchased from Red Hat
  - OpenAFS: no problem; sound/special video: additional effort

SL 4 and on: Changes

- SL3 is our first Linux ever with many years of support
- => SL4 was the occasion to make a few major changes
  - which is also one of the reasons why it's available so late
  - many months to get used to SL4/5
- no more HEPiX11 - incl. fvwm2
  - available: GNOME, KDE
  - lightweight window managers: IceWM, WindowMaker
- scrubbed a few legacy applications (plan,...)
- changes under the hood (profiles,...), hopefully not visible

What hasn't changed

- scientific software equipment
  - cernlib, root, maple, mathematica, ...
- browsers, mail readers, document viewers
  - firefox is the recommended browser
    - flash & java plugins, ...
  - pine still is the recommended mail reader
    - thunderbird available as is
  - gv, ... still available
- LANG is still set to C
  - we tried UTF-8,but it's a can of worms
  - users can set LANG in ~/.i18n if desired

AFS Sysnames

- primary sysname:
  - SL4
    - 32-bit: i386_linux26
    - 64-bit: amd64_linux26
    - these are the default sysnames as defined by the OpenAFS project
      - HH: i586_rhel40, amd64_rhel40, default sysnames are last in list
  - SL5
    - 32-bit: i586_rhel50
    - 64-bit: amd64_rhel50
- rest of sysname list:
  - primary sysnames of previous releases (down to DL5)
  - 64 -> 32 (amd64_rhel50, i586_rhel50, amd64_linux26, ...)

Why Users would want SL4/5

- responsiveness during I/O
  - SL3 is abysmal in this respect
    - even though performance is actually ok
    - we made serious efforts to improve this
      - to no avail
- more recent KDE/GNOME
- more recent gcc
  - SL4: 3.4.3
  - SL5: 4.1.1
    - g77 -> gfortran

Why Admins would want SL4/5

- the more exciting changes are under the hood:
  - security enhancements
    - Security Enhanced Linux ("SELinux")
      - initial release with SL4
      - major enhancements, modularization with SL5
    - Position Independent Executables (PIE)
    - common objective: make buffer overflows a non-issue
      - together with ExecShield (introduced with SL3)
    - should be invisible to users
    - but: steep learning curve for admins
  - virtualization
    - SL5 will come with Xen
    - has been driving (or slowing down) RHEL5 schedule

64-bit

- it's the future!
- farm will generally run 64-bit OS
  - with 32-bit runtime compatibility
    - all centrally provided libraries
  - standard for 2 years now
  - contact uco if your application requires a 32-bit environment
    - remaining 32-bit nodes will vanish eventually, or have restrictions
- 64-bit interactive systems available
  - for 2 years as well
- 64-bit desktops are possible with SL4 & 5
  - requires Dell Precision 370 or later

SL4: Status

- available now
- public preview systems:
  - sl4.ifh.de
  - sl4-64.ifh.de
- requires 6 GB root filesystem
  - 8 GB is better (and probabaly required for SL5)
  - more software installed locally
- User Information available in our Wiki:
  - http://dvinfo.ifh.de/SL4_User_Information

SL5: Status

- RHEL5 not yet released (ETA: "early in 2007")
- SL5 has to follow
- integration in Zeuthen well advanced:
  - started working with FC6, now working with EL5beta2
    - automatic installation/maintenance finished
    - most problems should be known and are being worked on
  - most scientific software is still missing
  - no user accessible preview systems yet
    - will be made available as soon as SL5 alpha/beta released
- ETA for a fully usable SL5 in Zeuthen: Q1/07

Timetable

- today SL4 available
- Q1/2007 SL5 available
- Q3/2007 next hardware generation, will no longer run SL3
- 10/2007 end of SL3 support by FNAL/CERN
- 10/2008 end of SL4 support (may be prolonged, though)
- 10/2010 end of support for RHEL3/CentOS3

Proposal

- skip SL4 where possible
  - aged already
  - has just a year longer to live than SL3
  - problem: ATLAS will probably need it for a while
    - CERN/LHC is locked on SL4 for LHC startup
- migrate farm, pubs, ... to SL5 in spring
  - will be able to run SL4 executables
  - and, hopefully, SL3
- get rid of SL3 desktops by 10/07

- opinions?