



# Speicherkonzepte

Ratgeber Verfügbarkeit, Management und Performance

transtec technologie kompass

**transtec**

# Verfügbarkeit, Management und Performance

sind die Schlagwörter bei der Konzeption moderner Speicherinfrastrukturen. Der transtec technologie kompass soll Ihnen bei der Planung Ihrer Speicherumgebung als Wegweiser mit praxiserprobten Empfehlungen zur Seite stehen.

Der Aufbau dieser Broschüre ist in die drei Hauptkapitel **Verfügbarkeit, Management** und **Performance** unterteilt und wird durch die Themengebiete **Festplattentechnologien** und Anwendungsperformance abgerundet.

Ein transtec Kundenprojekt beim Klinikum St. Marien Amberg zieht sich wie ein roter Faden durch die Broschüre. Dieses mit 554 Planbetten ausgewiesene Klinikum investierte im Jahre 2005 in eine PACS-Lösung – ein weiterer großer Schritt in die Digitalisierung der Radiologie.

Zur Speicherung der Bilddaten implementierte transtec eine hochverfügbare Storagevirtualisierungslösung. Diese Lösung verfügt über eine Kapazität von 2x 9 TByte gespiegelt und lässt sich um mindestens den Faktor 5 skalieren. Die in der Broschüre enthaltenen Projektbeispiele beziehen sich auf diese Lösung und beschreiben die Anforderungen und Lösungskomponenten in den Bereichen Verfügbarkeit, Management und Performance.



## Verfügbarkeit

- Systemverfügbarkeit
- Hochverfügbarkeit mittels Clusterlösungen
- Datenverfügbarkeit mittels Spiegelung/Replikation und Snapshots

Seite 4

## Management

- Management von Datenstrukturen
- Management von Speichersystemen
- Betriebsüberwachung und Ereignismeldung
- Storagevirtualisierung
- Hierarchisches Speichermanagement

Seite 6

## Performance

- Leistungsfaktor Speichersystem
- Leistungsfaktor Speichernetzwerk
- Skalierbarkeit von Performance

Seite 8

## Festplattentechnologien

Seite 10

## Anwendungsperformance

Seite 11

## Darstellung Speicherinfrastruktur Klinikum Amberg

Seite 12

## Weiterführende Informationen

Seite 13

## Kontakt und Impressum

Seite 14

# Verfügbarkeit

Heutige Unternehmensprozesse werden weitgehend durch IT unterstützt. Ob in der Entwicklung, Produktion oder Logistik, in Vertrieb und Marketing, Vertragsabwicklung und Lohnbuchhaltung bis hin zum Finanzwesen - die Abhängigkeit von diesen Daten und den entsprechenden Systemen und damit die Ansprüche an ihre Verfügbarkeit nehmen stark zu.

Damit sich Maßnahmen zur Erhöhung der Verfügbarkeit treffen lassen, werden die Bereiche der Systeme, Anwendungen und Daten jeweils gesondert betrachtet.



## Systemverfügbarkeit

Jede Komponente eines Systems, welche bei einem Ausfall zu einem Systemstillstand führen kann, wird als „Single Point of Failure“ bezeichnet. Zu diesen Komponenten zählen Prozessor, Speicher, Netzteil, Lüfter, Festplatte, Netzwerkkarte, Mainboard und Festplattencontroller. Durch die redundante Auslegung von Komponenten, die im Fehlerfall automatisch die Aufgaben der fehlerhaften Komponente übernehmen, lassen sich diese reduzieren; somit erhöht sich die Verfügbarkeit des Gesamtsystems. Der Austausch der fehlerhaften Komponente sollte dabei im laufenden Betrieb (Hotswap-Funktionalität) möglich sein, da die Ausfallzeit ansonsten zunächst nur abgefangen und verschoben, jedoch nicht behoben wäre.

Um die Verfügbarkeit von Daten auf Festplatten zu gewährleisten, lassen sich Festplatten mittels RAID-Controllern und deren Intelligenz in verschiedenen RAID-Levels konfigurieren. Je nach gewähltem RAID-Level werden die Festplatten gespiegelt bzw. so konfiguriert, dass der Ausfall einer Festplatte im Verbund nicht zum Verlust der Daten führt. Die verschiedenen Festplattentechnologien und deren Einfluss auf die Verfügbarkeit von Speichersystemen wird auf S. 10 gesondert betrachtet.

## Hochverfügbarkeit mittels Clusterlösungen

Sollen komplette Systeme sowie deren Anwendungen gegen Ausfall abgesichert werden, kann dies über den Einsatz von Clustern erreicht werden. Mit Hilfe eines HA-Clusters (HA = High Availability) lässt sich auch der Fehlerfall des Mainboards absichern, was innerhalb eines Einzelsystems nicht zu gewährleisten ist. Der HA-Cluster hält sowohl die Einzelsysteme als auch die Applikationen redundant vor und automatisiert die Übernahme der Funktionalität des Primärsystems durch das sekundäre System im Fehlerfall. HA-Cluster werden in der Regel bei Servern eingesetzt, bei denen sich der Datenbestand häufig ändert. Die Verbundsysteme greifen dabei auf einen gemeinsamen externen Datenbestand zu,

da die Daten für beide Systeme verfügbar sein müssen. Dieser befindet sich meist auf einem externen RAID-Plattensystem. Der Systemstatus der beteiligten Rechner wird regelmäßig gesichert, so dass bei Ausfall eines Servers ein anderer die Funktion übernehmen kann.

## Aktiv-Passiv oder Aktiv-Aktiv Cluster

Ein Aktiv-Passiv Cluster ist so konfiguriert, dass eines der beiden Systeme permanent produktiv ist, das andere jedoch nur im Falle eines Systemstillstands die Aufgaben des fehlerhaften Systems übernimmt. Beim Aktiv-Aktiv Cluster betreuen beide Systeme gleichzeitig verschiedene Bereiche, im Fehlerfall werden die Aufgaben des fehlerhaften Systems durch das Zweite mitübernommen. Dies setzt jedoch entsprechend freie Kapazitätsressourcen voraus.

Solche Clustersysteme lassen sich beispielsweise mittels des Microsoft® Cluster Service (MSCS) des Windows® Server 2003 konfigurieren. Auch im Linux-Umfeld sind verschiedene Clusterlösungen erhältlich.

## Datenverfügbarkeit mittels Spiegelung/Replikation und Snapshots

Neben der Verfügbarkeit von Server- und Speichersystemen gilt es auch die Verfügbarkeit der Daten durch den Aufbau von Redundanzen zu gewährleisten. Dies bedeutet, dass Daten über Spiegelungen/Replikationen synchron oder asynchron auf einem zweiten Speichersystem, das sich meist an einem anderen Ort befindet, redundant vorgehalten werden. Spiegelung und Replikation arbeiten zwar beide auf Blockebene, jedoch werden bei der Spiegelung logische Laufwerke gesichert, während bei der Replikation bestimmte Verzeichnisse oder Dateien als Quelle bestimmt werden. Im Folgenden steht der Begriff Spiegelung auch synonym für Replikation.

Bei der synchronen Datenspiegelung verfügen beide Systeme zu jedem Zeitpunkt über den selben Datenbestand. Das Netzwerk muss bei diesem Verfahren jedoch entsprechend ausgelegt sein.

Die synchrone Spiegelung hat zunächst den Vorteil, dass zu jeder Zeit eine identische Kopie aller Daten verfügbar ist. Falls Daten jedoch versehentlich geändert bzw. gelöscht oder durch Virusbefall zerstört werden, werden auch diese Änderungen zeitgleich auf dem redundanten System vollzogen. Im Falle der asynchronen Datenspiegelung, bei der jeweils zu einem festgelegten Zeitpunkt die aktuell vorhandenen Daten gespiegelt werden, hat der Administrator zumindest noch die Möglichkeit zwischen den Synchronisationsvorgängen entsprechende Maßnahmen einzuleiten. Erst im Anschluss daran werden die Daten an das zweite Plattensystem übermittelt. Eine asynchrone Spiegelung kann beispielsweise über das Asynchrones IP Mirroring (AIM) Modul der Firma DataCore™ realisiert werden. Beide Verfahren setzen voraus, dass die Speicherkapazität des redundanten Systems mindestens der Menge an Daten des Primärsystems entspricht.

Eine weitere Möglichkeit, um einen Datenbestand zu einem bestimmten Zeitpunkt zu sichern, bietet die Snapshot-Funktion.

Hier wird zu bestimmten Zeitpunkten ein Abbild (Momentaufnahme) der Datenblöcke abgelegt. Der Inhalt eines Datenblocks wird erst dann in den Speicher geschrieben, wenn dieser durch einen Schreibzugriff geändert werden soll. Vor Überschreiben des Datenblocks wird der aktuelle Stand dieses Blockes auf ein Sekundärsystem geschrieben, der bei Bedarf wieder auf das Primärsystem zurückgeschrieben werden kann. Mit Hilfe von Snapshots lassen sich somit durch den Benutzer verursachte logische Fehler rückgängig machen. Der Einsatz ist beispielsweise für Test- und Produktivumgebungen sehr sinnvoll. Zudem wird gegenüber klassischen Spiegelungen durchschnittlich nur etwa 20 - 30 % der Speicherkapazität benötigt.

	Anwendungsgebiete	Vorteile	Einschränkungen
<b>Synchrone Spiegelung/ Replikation</b>	Applikationen, die nicht geschwindigkeitskritisch sind, bei denen es jedoch auf eine hundertprozentige Vollständigkeit der Daten ankommt	Zu jedem Zeitpunkt exakte Kopie des lokalen Plattensystems durch die synchrone Spiegelung/ Replikation	Automatische Übernahme des Zustands auf den Spiegel im Falle der Inkonsistenz der Datenbank aufgrund eines logischen Fehlers
	Entfernungen < 100 km		
<b>Asynchrone Spiegelung/ Replikation</b>	Spiegelung/Replikation über größere Distanzen zwischen Rechenzentren	Kein Effekt der eingeschränkten Performance im Vgl. zur synchronen Spiegelung/Replikation	Risiko möglicher (minimaler) Verluste bei der Rekonstruktion von Daten durch die zeitversetzte Spiegelung/Replikation
	Anbindung von Außenstellen zur Spiegelung/Replikation der Daten in das Rechenzentrum für das Backup		
	Entfernungen > 100 km	Übermittlung des Acknowledge vor der Spiegelung/ Replikation an den Application Server (synchrone Variante: Übermittlung nach dem Schreiben auf das Sekundärsystem)	
<b>Snapshots</b>	Stündliches Backup	Wiederherstellung von Datenzuständen unterschiedlicher Zeitpunkte	Benötigter Speicherplatz entspricht u. U. dem des Originals, z. B. bei Testumgebungen und Daten, die sich ständig ändern.
	Einsatz für Test- und Entwicklungsumgebungen, z. B. zum Test neuer Datenbankversionen	Schnelle Wiederherstellung möglich	



## Projektbeispiel Klinikum Amberg

### Anforderungen an die Verfügbarkeit der Speicherumgebung:

Sicherstellung der ständigen Zugriffsmöglichkeiten auf die radiologischen Unterlagen und Patientendaten durch multiple Datenbevorratung, hochverfügbare Speichersysteme und ein redundantes Netzwerk.

### Lösungskomponenten:

- 2 Storage Domain Server (SDS) zur Storageverwaltung im Clusterbetrieb mittels DataCore™ SANmelody Software. Bestückung der SDS-Server mit redundant ausgelegten Fibre Channel Host Bus Adapters, gespiegelten Systemplatten und redundanten Netzteilen.
- 2 RAID-Systeme mit Anbindung an die SDS-Server zur Datenspiegelung. Redundante Auslegung der Netzteile und RAID-Controller sowie Einsatz von RAID-Levels zur Absicherung eines Festplattenausfalls.
- Redundante Auslegung der Fabrics
- Redundante Anbindung der Application Server

Der schematische Aufbau dieser Lösung ist auf S. 12 abgebildet.

# Management

IT-Administratoren haben eine Vielzahl an unterschiedlichen Aufgaben im Bereich des Speichermanagements zu erfüllen.

Mit Zunahme der Komplexität dieser Aufgaben steigt auch der Bedarf an Zeit und Schulungen. Das Bestreben sollte daher sein, die Speicherumgebung so auszulegen, dass selbst mit zunehmendem Datenaufkommen der Administrationsaufwand gesenkt werden kann.

Des Weiteren sollte die Speicherinfrastruktur skalierbar, leistungsfähig und je nach Einsatzgebiet auch hochverfügbar sein. Bei einer geplanten Umstrukturierung sollte das entsprechende Konsolidierungspotential der vorhandenen Umgebung analysiert und darauf geachtet werden, dass sich bestehende Speichersysteme in die neue Umgebung integrieren lassen - sofern gewünscht.

## Management von Datenstrukturen

Die Tatsache, dass der Bedarf an Speicherkapazität stetig wächst, erfordert eine kontinuierliche Analyse der vorhandenen Speicherumgebung. Die folgenden Fragen stehen dabei im Vordergrund:

- Wie werden die vorhandenen Kapazitäten verwendet?
- An welchen Stellen im SAN befindet sich der größte Bandbreitenbedarf?
- Wie sieht der historische Speicherbedarf aus?

Diese und weitere Aspekte lassen sich durch den Einsatz von Storage Management Software ermitteln. transtec bietet in Kombination mit SANmaestro der Firma DataCore™ eine Virtualisierungslösung an, die Analyse und Monitoring Tools für diesen Zweck bereitstellt.

## Management von Speichersystemen

Die Verwaltung von RAID-Systemen kann über unterschiedliche Oberflächen erfolgen. transtec PROVIGO 600 Stagesysteme bieten die Möglichkeit, seriell via RS-232C oder über Ethernet via Telnet auf ein Firmware-embedded Managementtool zuzugreifen. Die ASCII-Oberfläche eignet sich für die Verwaltung und Konfiguration des RAID-Systems, falls ein Unix/Linux Host zur Administration verwendet wird.

Zur grafischen Unterstützung der Verwaltung kommt das JAVA Tool RAIDConn zum Einsatz, welches erweiterte Informationen zum Systemzustand liefert und die Speicherung des NVRAMS in eine Datei erlaubt. Ein zentrales Management von mehreren transtec PROVIGO 600 Storage-Systemen ist über die RAIDConn Software ebenfalls möglich. Hierfür werden verschiedene Profile in der Software konfiguriert, die es erlauben, zwischen unterschiedlichen RAID-Systemen zu wechseln.

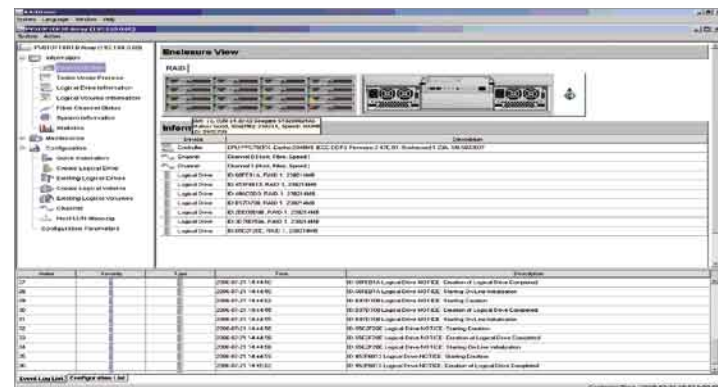
## Betriebsüberwachung und Ereignismeldung

Die Zeiten, in denen am Host-System ein Software Agent zur RAID-Überwachung notwendig war, sind vorbei. Moderne RAID-Systeme verfügen über ein eigenständiges, ausgeklügeltes Alarmsystem. Zu den integrierten Überwachungsdiensten zählen u. a. die SCSI Enclosure Services (SES), die Parameter wie Temperatur, Lüfterdrehzahl oder Laufwerksausfälle über den FC- oder SCSI-Kanal an das RAID-Betriebssystem melden. Die Ereignisse lassen sich in drei Klassen einordnen: Information, Warnung und Fehler. Das Betriebssystem wiederum kann ein aufgetretenes Ereignis dem Administrator über unterschiedliche Mechanismen melden. Folgende Benachrichtigungsoptionen werden von transtec Systemen unterstützt:

- Email
- SNMP Traps
- LAN Broadcast
- Pager

## Storagevirtualisierung

Das Problem des kontinuierlich zunehmenden Speicherbedarfs, die Herausforderung der Zentralisierung des Speichermanagements und die flexible



transtec RAID Management Software RAIDConn

Aufteilung der Ressourcen über verschiedene I/O-Kanäle ist mit einer Virtualisierungslösung zu bewältigen. transtec bietet mit der DataCore™ Software SANmelody und SANSymphony zwei Virtualisierungslösungen, die dem Anspruch eines Enterprise Storage Managements entsprechen. Die Lösung bietet u. a. die Möglichkeit vorhandene Speicherkapazitäten zu überbuchen, gespiegelte Volumes und Snapshots zu generieren, asynchrone IP-Spiegel aufzubauen und dabei vorhandene RAID-Systeme weiter verwenden zu können. Sie kann sowohl im Unix, Linux als auch Microsoft® Umfeld eingesetzt werden und ist somit auch für heterogene Infrastrukturen geeignet.

Bei dieser Virtualisierungslösung sind die RAID-Systeme nicht wie üblich direkt mit den Hostsystemen verbunden, da als weitere Abstraktionsschicht ein sogenannter Storage Domain Server (SDS) zwischengeschaltet wird. Die Anbindung der vorhandenen bzw. neuen RAID-Systeme findet über SCSI oder FC an den SDS statt. Die verfügbaren Kapazitäten werden durch den SDS zu Storagepools zusammengefasst, aus denen schließlich Volumes generiert werden, die wiederum über iSCSI oder FC an die Hostsysteme gemeldet werden.

Auf S. 12 befindet sich eine schematische Darstellung dieser Lösung.

Da der SDS-Server die Ausgabe der Speicherkapazität verwaltet, können erweiterte Funktionen genutzt werden.

#### **Virtuelle Kapazität / Auto Provisioning:**

Von Überbuchung spricht man, wenn die Speicherkapazität, die dem Host insgesamt gemeldet wird, den tatsächlichen physikalisch vorhandenen Speicher übersteigt. Besitzt z. B. der SDS-Server in seinen Storagepools 5 TByte, so kann er über die Auto Provisioning Funktion dennoch an fünf Hosts gleichzeitig je 2 TByte melden. Somit ist es möglich, wachsende Strukturen zu berücksichtigen und unmittelbar 2 TByte zur Verfügung zu stellen, auch wenn aktuell nur 500 GByte in Gebrauch sind. Der SDS-Server belegt in den Storagepools jedoch nur den tatsächlich verwendeten Speicherbedarf und meldet sich, sobald eine Grenzmarke überschritten wird.

**I/O-Handling:** Der SDS-Server sammelt die eingehenden I/O-Anforderungen und schreibt sie im Load Balancing Modus auf die RAID-Systeme. Dadurch lässt sich ein wesentlich besserer Durchsatz erreichen, was zu einer optimalen Auslastung der RAID-Systeme führt.

### **Hierarchisches Speichermanagement**

Unter hierarchischem Speichermanagement - auch bekannt als HSM, Tiered Storage oder ILM - versteht man die Klassifizierung von Daten und deren Zuordnung zu definierten Speicherklassen. Man verspricht sich hiermit eine höhere Effizienz in punkto Kosten, Performance und Verfügbarkeit.

Anwendungen und Serverdienste haben unterschiedliche Anforderungen bezüglich Zugriffszeit und Verfügbarkeit an die Speicherrressourcen. Die Home-Verzeichnisse der Mitarbeiter lassen sich beispielsweise auf RAID-Systemen mit kostengünstigen SATA-Datenträger speichern. Unternehmenskritische Dienste wie SQL Server hingegen lagern ihre Daten auf sehr leistungsfähigen und hochverfügbaren Systemen. Dadurch entwickeln sich aus der Anwendungsart heraus unterschiedliche Speicherklassen, die separat voneinander geplant werden müssen. Auf S. 10 werden Empfehlungen gegeben, welche Festplattentechnologie für welche Anwendungsbereiche geeignet ist.

SATA- und SAS-Festplatten sind prinzipiell für unterschiedliche Einsatzgebiete konzipiert. Attraktiv ist jedoch die parallele Einsatzmöglichkeit innerhalb eines Systems. Alle SAS-Festplattencontroller erkennen automatisch die installierten Festplattentypen und unterstützen beide Technologien. Dies bietet einen großen Vorteil beim Aufbau von gestuften Speicherarchitekturen, da eine modulare Trennung zwischen Primär- und Sekundärspeichern innerhalb eines Systems möglich ist und sich dadurch Kosten der Gesamtlösung sparen lassen.



## Projektbeispiel Klinikum Amberg

### **Anforderungen an das Management der Speicherumgebung:**

Einfache Verwaltung und Erweiterbarkeit der Speicherumgebung sowie Fernwartung durch das transtec Competence Center.

### **Lösungskomponenten:**

- Virtuelle Volumes mit 2 TByte Kapazität zur Versorgung der Server
- Bereitstellung der Volumes über eine redundante Storagevirtualisierungslösung
- Direkte Fehlermeldungen der RAID-Systeme, SDS-Server und FC-Switches per Email an den transtec Support
- WEB-based Management Funktionalität der FC-RAID-Systeme und FC-Switches für eine einfache Administration und Monitoring
- Dedizierter VPN-Tunnel für den transtec Support zur Systemdiagnose und -überwachung
- Zugriff auf Storage Server per RDP-Protokoll
- 24 x 7 x 4 Service

# Performance

Durch Speicherkonsolidierung und die Digitalisierung von Geschäftsprozessen wachsen die Leistungsanforderungen an die zentrale Speicherlösung. Doch welche Faktoren beeinflussen die System- und Anwendungsperformance und worauf ist zu achten?

Das zentrale Kriterium jenseits aller Werbung bleibt die Physik des Speichermediums Festplatte. Um eine bestimmte Bandbreite oder OLTP-Transaktionsrate mit akzeptabler Latency zu erzielen, muss unabwendbar eine abschätzbare Mindestanzahl Festplatten eingesetzt werden.

Diese und weitere Aspekte zu Festplatten werden in den Abschnitten Festplattentechnologien und Anwendungsperformance auf den Seiten 10 und 11 gesondert betrachtet.

## Leistungsfaktor Speichersystem

Während im Serverumfeld die Leistungsfähigkeit des zentralen Prozessors und die interne PCI-Busarchitektur weithin beachtete Entscheidungskriterien sind, ist dies im Speicherumfeld zu Unrecht seltener der Fall. Dabei liegen hier Leistungsunterschiede bis zu 50 % begründet. Insbesondere bei skalierenden Speichersystemen ist der Controller oft ein Flaschenhals.

In Speichersystemen der Klasse unter 25.000 € werden die Embedded Prozessorfamilien Intel® XScale®/IOP und IBM® PowerPC® eingesetzt. Und wie bei Servern haben die RISC-CPU's auch hier zumeist einen deutlichen Leistungsvorsprung vor ihren preislich günstigeren Pendanten auf x86-Basis. Relevant ist zudem die eingesetzte CPU-Generation mit Kenngrößen wie Taktung, L2-Cache, SDRAM-Typus und vor allem Art und Anzahl der PCI-Busse für I/O-Prozesse. Die maximal erzielbare Leistung entspricht der halben PCI-Busbreite. Für rechenintensive RAID-6 Paritätskalkulationen sollten zudem spezielle ASICs zur Entlastung der Embedded CPU vorhanden sein.

Die Bedeutung der absoluten Größe des Caches in Speichersystemen wird hingegen eher überschätzt (siehe Abbildung 1). Die Ausnahme hiervon sind Datenbanksysteme, die erlauben, Tablespace und

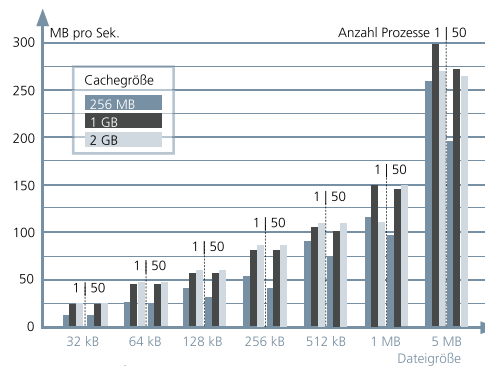


Abb. 1: Performance Messung mit Iometer (50 % Random Read/Write)

Logfiles komplett im schnellen Cache zu belassen oder Hot Zones gezielt in selbigen auszulagern. Jedoch muss hierzu auf den Cache des Speichersystems wie auf eine virtuelle Systemplatte geschrieben werden können. Eine Fähigkeit, welche primär Enterprise Disk Arrays wie HDS TagmaStore® USP oder EMC® Symmetrix® aufweisen. Eine Alternative ist der Einsatz von Solid State Disks (SSD) zur Auflösung von Hot Spots, welche preislich im unteren fünfstelligen Bereich liegen.

Wichtiger als die Größe ist eine optimale Nutzung des Caches durch den Speichercontroller. Hier finden sich erhebliche Unterschiede, welche Leistungsunterschiede bis über 20 % bedingen können. Ohne Benchmarks ist dies aber nur schwer beurteilbar. Positive Indikatoren sind jedoch Fähigkeiten wie partitionierbarer und gezielt allozierbarer Cache, dedizierte Cache-Policies pro Volume, dynamische Anpassung an aktuelle I/O-Muster oder Best Practice Voreinstellungen für Anwendungsprofile wie OLTP-Betrieb oder Audio-/Video-Streaming.

## Leistungsfaktor Speichernetzwerk

Der klassische Direktanschluss an bis zu vier Server ist für viele KMUs immer noch die günstigste Lösung. Statt SCSI sollte möglichst ein netzwerkfähiger Host-Anschluss gewählt werden. Sobald verfügbar empfiehlt sich hierfür SAS aufgrund vergleichbarer Kosten bei einer deutlich höheren Bandbreite von bis zu 1200 MB/s pro 4-fach-Anschlusskabel. Auch ist mit SAS ein auf einen 19"-Schrank beschränktes Mini-SAN aufbaubar.

Fibre Channel ist die primäre Wahl für Speichernetzwerke und schlägt mit 400 MB/s Bandbreite pro Kanal bei unter 6 % Protokoll-Overhead sowohl SCSI/SAS als auch Ethernet. Dem stehen hohe Infrastrukturkosten und begrenzte Reichweite für Ausfallrechenzentren entgegen. Hierin liegen die Stärken von 1 Gbit-iSCSI. In der Bandbreite eingeschränkt, bringt iSCSI im Praxiseinsatz jedoch erstaunlicherweise kaum Nachteile in Bezug auf die Latency (siehe Abbildung 2). Mit ausreichenden GbE-Leitungen, LAN-Switches mit gutem Trunking-Support und dem Einsatz von iSCSI HBAs oder TOE-

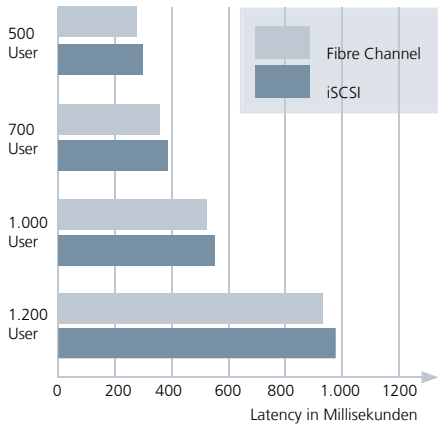


Abb. 2: Vergleich der Latency im SAN für MS Exchange Server 2003

Karten ist durchaus der Aufbau hochperformanter IP-SANs möglich. Die Belastung einer Intel® Xeon® 3.60 GHz CPU für zwei GbE-Ports ohne spezielle Offload-Netzwerkadapter liegt bei ca. 25 %, so dass für weniger belastete Server auch ein Software iSCSI Initiator einsetzbar ist. Für Dual-Core-Server-systeme mit Intel® I/OAT-Unterstützung sinkt die Belastung sogar auf unter 10 %.

Sind NAS-Systeme eine weitere Option für Applikationen, die neben RAID-Systemen auch über Netzwerkprotokolle wie NFS betrieben werden können? Schreibintensive Vorgänge wie profitieren von iSCSI, da NFS nur begrenzt asynchrone Schreibvorgänge im Client unterstützt, was zu merklichen Leistungseinbußen führt. Auch für Metadaten-intensive Anwendungen ist iSCSI in TPC-Benchmarks gegenüber NFS bis zu doppelt so schnell, wofür in erster Linie aggressives Metadaten-Caching und die asynchrone Aggregation von Metadaten-Updates unter iSCSI verantwortlich zeichnen. Im Praxiseinsatz mit Datenbank-Systemen wie Oracle® reduzieren sich die Unterschiede zwischen NFS und iSCSI jedoch bei Nutzung aller Tuningmöglichkeiten auf den einstelligen Prozentbereich (siehe Abbildung 3).

## Skalierbarkeit von Performance

Für Höchstleistungsanforderungen von mehreren 10.000 I/Os oder Gigabyte-Bandbreiten sind zwingend multiple Speichercontroller erforderlich. Herausragend sind monolithische Enterprise Disk Arrays, die über bis zu 64 Prozessoren mit eigenem Cache und I/O-Bussen verfügen, jedoch auch siebenstellige Investitionen erfordern. Eine günstigere Möglichkeit ist der Einsatz mehrerer günstiger Speichersysteme und eine Zusammenfassung der bereitgestellten Volumes auf Betriebssystemebene. Dies ist jedoch sehr aufwändig und bietet keine Möglichkeiten zur aktiven Lastverteilung. Diese kann eine Speichervirtualisierungsschicht bieten, welche vom konkreten System abstrahiert und erhebliche Vorteile auch bei der Allokation und Administration des Speichers bietet.

Die interessanteste und neueste Alternative sind Storage-Cluster. Analog zu HPC-Clustern werden hier zahlreiche kleinere Speicherknoten zu einem massiv parallelen Speichersystem zusammenschaltet. Die Gesamtleistung konkurriert durch die hohe Anzahl an Speichercontrollern, Cache und Hostleitungen mit Enterprise Disk Arrays, ist jedoch in Anschaffung und Wartung erheblich günstiger.

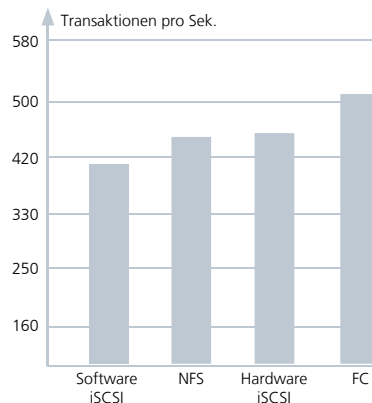


Abb. 3: Oracle® 10g Performancevergleich mit transtec Unified Storage



## Projektbeispiel Klinikum Amberg

### Anforderungen an die Performance der Speicherumgebung:

Schnelle Speicherung und Wiedergabe der radiologischen Bilddaten des PACS<sup>1</sup>-Systems sowie Migration der Daten der elektronischen Patientenakten auf untenstehende Storagelösung infolge der hohen I/O-Leistung.

### Lösungskomponenten:

- SDS-Server mit einer theoretischen Leistung von bis zu 220.000 I/O pro Sek.
- Dual-Controller RAID-Systeme, die eine Lastverteilung der RAID-Sets erlauben
- Redundante 2 Gbit FC-Anbindung (400 MB/s FD) der Server
- Strukturierung der Datenbestände in unterschiedliche Storagepools
- Zweistufiges Caching der I/O-Requests durch SDS-Server und darunterliegende RAID-Systeme
- Optimierung der RAID-Systeme für Random I/O
- Load Balancing der I/O-Pfade vom SDS-Server zum RAID-System

<sup>1</sup> Picture Archiving and Communication System

# Festplattentechnologien

Festplatten sind das zentrale Kriterium, wenn man die Verfügbarkeit und Performance von Speichersystemen betrachtet. Je nach Anforderungen der Anwendungen und verfügbarem Budget fällt die Auswahl auf SATA-, SCSI-, SAS- oder FC-Festplatten.

## Performance

SATA-Festplatten sind die optimale Wahl für den Transfer großer Dateien (> 100 KB). Jedoch arbeiten Datenbanken wie beispielsweise Exchange, Oracle oder SQL zumeist mit Datenpaketen von 512 Byte bis 8 kByte. Hier ist SATA erheblichen Leistungsbeschränkungen unterworfen. Wir empfehlen nur kleinere OLTP-Systeme mit bis ca. 100 Usern auf einem SATA-RAID zu betreiben. Ansonsten sind SCSI, SAS oder Fibre Channel die bessere Wahl. Der Grund liegt zum einen in der höheren Umdrehungszahl, wodurch Zugriffe schneller erfolgen. Zum anderen können Befehle besser vorsortiert werden, um zahlreiche kleine Zugriffe schneller abarbeiten zu können. SATA-2 hat eine maximale NCQ-Sortiertiefe von 32, während SCSI, SAS und FC bei 256 liegen.

Generell empfiehlt transtec den Einsatz von seriellen Architekturen wie SATA, SAS und FC, da hier keine Performance-Engpässe wie bei einer über mehrere Festplatten geteilten Kanalbandbreite auftreten.

## Zuverlässigkeit

Alle von transtec eingesetzten Festplatten sind für den 24x7 Dauerbetrieb spezifiziert. Jedoch gibt es bei SATA- gegenüber Enterprise-Festplatten Nachteile in Bezug auf Duty Cycle, Temperatur und Vibrationstoleranz. SATA-Festplatten sind in der Regel lediglich für aktive Zugriffe während 10 - 20 % der Betriebszeit ausgelegt. Finden ständig I/O-Zugriffe statt, erhöht sich die Ausfallrate signifikant. Die Betriebstemperatur sollte 30 °C nicht überschreiten. Der gelistete MTBF stellt einen statistischen Wert dar, der sich bei Veränderung der Parameter Duty Cycle oder Temperatur ebenfalls verändert.

	PATA	SATA	SCSI	SAS	FC
<b>Typ</b>	Parallel	Seriell	Parallel	Seriell	Seriell
<b>Bandbreite pro Kanal</b>	133 MB/s	300 MB/s	320 MB/s	300 MB/s	200 / 400 MB/s
<b>Anzahl Festplatten pro Kanal</b>	max. 2	1	max. 14	1 (max. 128 mit Expander)	1
<b>Performance pro Festplatte</b>	≤ 60 MB/s	≤ 65 MB/s	≤ 90 - 105 MB/s	≤ 105 MB/s	≤ 95 - 110 MB/s
<b>Umdrehungsgeschwindigkeit</b>	5.400 - 7.200 U/min	7.200 U/min	10.000 - 15.000 U/min	15.000 U/min	10.000 - 15.000 U/min
<b>Organisation der Schreib-/Lesezugriffe</b>	-	Native Command Queuing (NCQ)	Tagged Command Queuing (TCQ)		
<b>Leistungsgrenze pro Kanal</b>	ab 2 Platten	-	ab 4 Platten	-	-
<b>Anwendungsgebiete</b>	Optimal für Transfer großer Dateien (> 100 KB) <ul style="list-style-type: none"> <li>• DBs bis ca. 100 Nutzer</li> <li>• A/V-Serving und -Streaming</li> <li>• Fileserver</li> <li>• Backup- und Archivsysteme</li> </ul>		Transaktionsintensive Anwendungen mit kleinen Datenblöcken <ul style="list-style-type: none"> <li>• OLTP-Datenbanksysteme</li> <li>• Mailingsysteme</li> </ul>		

	SATA	SCSI/SAS/FC	Nearline SATA (WD RAID Edition)
<b>MTBF<sup>1</sup></b>	1 Mio. Std.	1.4 Mio. Std.	1.2 Mio. Std.
<b>Power-on Hours (PoH)<sup>2</sup></b>	24 x 7 (8736 h)	24 x 7 (8736 h)	24 x 7 (8736 h)
<b>Duty Cycle<sup>3</sup></b>	10 - 20 %	80 - 100 %	80 - 100 %
<b>Temperatur</b>	25 °C	60 °C	60 °C
<b>Ausfallrate</b>	0,9 % (24x7) > 3 % (unter Höchstlast)	0,6 %	1,3 % (unter Höchstlast)
<b>Anwendungsgebiete</b>	Kostengünstiger Sekundärspeicher	<ul style="list-style-type: none"> <li>• Primärspeicher mit Dauerbetrieb</li> <li>• Cluster- &amp; RAID-Systeme</li> </ul>	

<sup>1</sup> MeanTime Between Failure (MTBF): Statistische Angabe, die sich auf hohe Plattenanzahlen bezieht. Bei beispielsweise 1000 SATA-Festplatten im Dauerbetrieb sind während eines Jahres (8736 Std.) durchschnittlich 9 Ausfälle zu erwarten

<sup>2</sup> Power-on Hours (PoH): Betriebsstunden, während der die Platte unter Strom steht.

<sup>3</sup> Duty Cycle: Prozentsatz der Betriebsdauer (Power-on Hours), während der aktive I/O-Zugriffe auf die Festplatte stattfinden.

# Anwendungsperformance

Unterschiedliche Anwendungen wie Online-Transaktionssysteme (OLTP), Data Warehousing oder sequentielle Filedaten erfordern eigene Methoden der Performanceoptimierung. In OLTP-Systemen ist Hardware oft weniger entscheidend als Datenbankdesign und Query-Optimierungen. Doch neben größerem Servercache profitieren DB-Administratoren auch von Speichercache-Tuning und einer optimalen Verteilung der Daten auf die Festplatten.

In OLTP-Systemen kommt es auf die I/O-Leistung an. Datenbanken sollten im Cache selbst oder zumindest auf schnelldrehenden Festplatten liegen. Zudem sollte der Speichercache für Schreiboperationen optimiert sein und jede Art von I/O-Aktivität (Datenbank, Transaktions-Logfiles, tempdb etc.) auf separaten Festplatten, RAID-Sets und idealerweise auch separaten Speichercontrollern arbeiten. Anwendungen mit lese- bzw. schreibintensiven Operationen (beispielsweise OLAP/OLTP) oder Anwendungen mit primär sequentiellen bzw. zufälligen Zugriffen sollten nicht auf dem selben Speichersystem liegen. Zumindest sollten sie jedoch auf getrennten Controllern und Cache-Partitionen untergebracht sein.

Auch der gewählte RAID-Level hat starke Auswirkungen auf die Leistung. Komplexe Parity-Level erfordern ausreichend leistungsstarke Controller sowie eine höhere Festplattenanzahl zur Kompensation des höheren I/O-Aufwandes für Schreiboperationen (siehe Tabelle).

Ein Microsoft® Exchange Mailbox Heavy User mit über 200 empfangenen oder gesendeten Emails pro Tag generiert bis zu zwei I/O-Zugriffe pro Sekunde, während der durchschnittliche User mit 100 Emails bei circa 0,75 I/O pro Sekunde liegt. Für tausend Mailboxen mit 900 normalen und 100 Heavy Usern wäre somit, bei einer Konfiguration mit RAID-Level 6, ein Speichersystem mit 2.450 I/Os respektive mindestens 25 Festplatten (10.000 U/min) erforderlich.

Für OLTP-Systeme mit hoher Last sind schnell-drehende Festplatten die bessere Wahl. Diese bieten eine um ca. 30 % höhere I/O-Leistung für 45 - 70 % Aufpreis. Auch für größere Datenbanken rentiert sich die Investition oft aufgrund der geringeren Ausgaben für hohe Skalierbarkeit oder zusätzliche Erweiterungsboxen. Im allgemeinen profitieren OLTP-Systeme stärker von der Investition in zusätzliche Festplatten als der Anschaffung zusätzlicher Speichercontroller. Für OLAP-Applikationen, Streaming und Backup-to-Disk ist meist die Bandbreite des Speichercontrollers der limitierende Faktor. Mehrere kleinere Speichersysteme und stärkere Speichercontroller sind die bessere Investition. Schließlich können Solid State Disks (SSD) für OLTP-Systeme mit Hot Spots und auch nach Plattenoptimierung anhaltend hoher I/O-Wait Time ein Ausweg sein. Anwender berichten von erheblichen Leistungssteigerungen bei ansonsten identischer Server- und Speicherhardware. Die Kosten sind jedoch ebenso erheblich.

Konfiguration	I/O pro Lesezugriff	I/O pro Schreibzugriff	I/O pro 100 OLTP-Zugriffe*	I/O-Effizienz*
RAID-0	1	1	100	100 %
RAID-1	1	2	130	77 %
RAID-0+1	1	2	130	77 %
RAID-5	1	4	190	52 %
RAID-6	1	7	280	35 %

\* basierend auf einem typischen OLTP-Mix aus 30 % Schreib- und 70 % Lese-Operationen

# Schematische Darstellung einer hochverfügbaren Storagevirtualisierungslösung

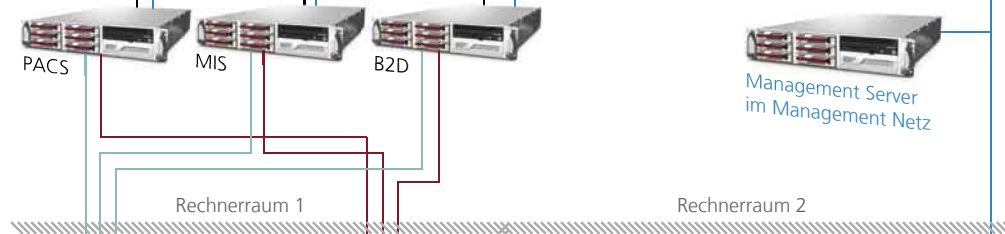
am Beispiel des Klinikums Amberg

Benutzer



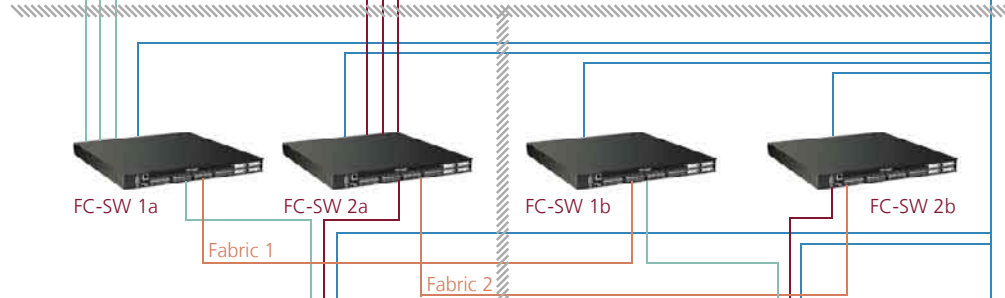
Applikationen

- PACS/MIS/B2D (Anbindung weiterer Datenbank-/Fileserver etc. möglich)



Anbindung an die Storage Domain Lösung

- Fibre Channel (auch iSCSI Anbindung möglich)

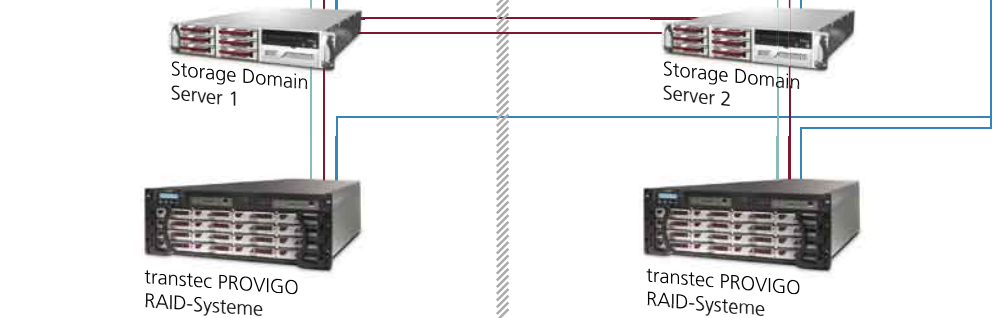


Storageverwaltung

- DataCore™ SANmelody

Storage Hardware

- transtec PROVIGO RAID-Systeme



PACS = Picture Archiving and Communication System  
MIS = Management Information System  
B2D = Backup-to-Disk  
LAN = Local Area Network  
FC-SW = Fibre Channel Switch

— = Management Netzwerk  
— = Primärer Datenpfad  
— = Sekundärer Datenpfad  
— = Fabric-Verbindung  
— = Local Area Network



## Weiterführende Informationen

### transtec Informationen zu Produkten, Lösungen und Technologien

transtec AG	Informationen und Angebote zu transtec Hardware	<a href="http://www.transtec.de">http://www.transtec.de</a>
transtec Competence Center	Informationen und Angebote zu transtec Lösungen im Speicher- und Serverumfeld	<a href="http://www.transtec-cluster.com">http://www.transtec-cluster.com</a>
transtec IT-Kompodium	Ausführliches IT-Technologie-Kompodium	<a href="http://www.transtec.de/go/kompodium">http://www.transtec.de/go/kompodium</a>
DataCore™	Hersteller von Softwareapplikationen im SAN-Umfeld	<a href="http://www.datacore.com">http://www.datacore.com</a>

### Webseiten mit Neuigkeiten zu Technologiethemem

tecCHANNEL	Plattform für Computertechnik und Betriebssysteme	<a href="http://www.tecchannel.de">http://www.tecchannel.de</a>
speicherguide.de	Plattform für Storage, Datenspeicherung und -management	<a href="http://www.speicherguide.de">http://www.speicherguide.de</a>

### Organisationen und Vereinigungen im Speicher- und Netzwerkumfeld

SNIA – Storage Networking Industry Association	Vereinigung zur Unterstützung der Industrie bei der Ausbildung und Standardisierung im Speicherumfeld	<a href="http://www.snia.org">http://www.snia.org</a>
Serial ATA International Organization	Vereinigung zur Unterstützung der Industrie bei der Implementierung und Spezifikation von SATA	<a href="http://www.serialata.org">http://www.serialata.org</a>
STA – SCSI Trade Association	Vereinigung zur Förderung der SCSI-Technologie	<a href="http://www.scsita.org">http://www.scsita.org</a>
FCIA - Fibre Channel Industry Association	Vereinigung zur Förderung der Fibre Channel Technologie	<a href="http://www.fibrechannel.org">http://www.fibrechannel.org</a>
IEEE - Institute of Electrical and Electronics Engineers	Vereinigung zur Förderung von Technologiethemem	<a href="http://www.ieee.org">http://www.ieee.org</a>
ASNP - Association of Storage Networking Professionals	Offenes Forum für Spezialisten im Speichernetzwerkumfeld	<a href="http://www.asnp.org">http://www.asnp.org</a>

#### **transtec AG**

Waldhörnlestrasse 18  
D-72072 Tübingen  
Tel.: +49 (0) 7071/703-400  
Fax: +49 (0) 7071/703-90 400  
transtec@transtec.de  
www.transtec.de

Projektleitung und Redaktion:  
Texte und Konzeption:

Grafische Gestaltung:

Weitere Informationen zu transtec Produkten und Lösungen finden Sie unter: [www.transtec.de](http://www.transtec.de) und [www.transtec-cluster.com](http://www.transtec-cluster.com)

#### **transtec Ges.m.b.H.**

Jedleseerstrasse 3/Top 11  
A-1210 Wien  
Tel.: +43 (0) 1/726 60 90 11  
Fax: +43 (0) 1/726 60 90 99  
transtec.at@transtec.at  
www.transtec.at

Markus Lohmüller, Marketing Manager | [Markus.Lohmueller@transtec.de](mailto:Markus.Lohmueller@transtec.de)  
Holger Hennig, Head of Competence Center | [Holger.Hennig@transtec.de](mailto:Holger.Hennig@transtec.de)  
Andrea Rudolf, Senior Storage Consultant | [Andrea.Rudorf@transtec.de](mailto:Andrea.Rudorf@transtec.de)  
Marco Poggioli, Product Manager Storage | [Marco.Poggioli@transtec.de](mailto:Marco.Poggioli@transtec.de)  
Sandra Kammerer, Foto- und Kommunikationsdesign | [sk@kammererkommunikation.de](mailto:sk@kammererkommunikation.de)

#### **transtec Computer AG**

Riedmattstrasse 9  
CH-8153 Rümlang  
Tel.: +41 (0) 44/818 47 00  
Fax: +41 (0) 44/818 47 20  
transtec.ch@transtec.ch  
www.transtec.ch



transtec technologie kompass

© transtec AG, September 2006

Alle verwendeten Abbildungen sind Eigentum der transtec AG und dürfen nur mit Genehmigung der transtec AG verwendet, vervielfältigt und veröffentlicht werden. Keine Haftung für fehlerhafte und unterbliebene Eintragungen.

**transtec**