



WESTFÄLISCHE
WILHELMS-UNIVERSITÄT
MÜNSTER

Warum kaufen wir Speicher nicht bei Saturn?

Technische Hintergründe der aktuellen Kaufentscheidung

CNMEMORY Airy externe Festplatte 3,5 Zoll 3 TB



CnMemory®
quality products from CHIPS and MORE

Testurteil: Computerbild Ausgabe 17/2011 //
Testergebnis: gut (2,25), Festplattentyp:
Extern, Anschlüsse: USB 3.0, Formfaktor
(Zoll): 3,5 Zoll, Formfaktor (cm): 8,89 cm,...

★★★★★ (1)

109,-

Saturn Preis

inkl. MwSt. zzgl. Versand
sofort lieferbar

Technische Merkmale	
✓ Festplattentyp:	Extern
✓ Anschlüsse:	USB 3.0
✓ Formfaktor (Zoll):	3,5 Zoll
✓ Formfaktor (cm):	8,89 cm
✓ Speicherkapazität:	3 TB
✓ Übertragungsrate:	bis zu 5 GBit/s, sofern USB 3.0 kompatible Geräte verwendet werden
✓ Unterstützte Betriebssysteme:	ab Windows 2000 und Mac OS 10 oder höher / 3 TB ab Windows 7
✓ Artikelnummer:	1415368
Ausstattung	
✓ Infodisplay:	LED Funktionsanzeige. Die LED ist im angeschlossenen Zustand blau und sonst weiß.
✓ Besondere Merkmale:	Sofort betriebsbereit, durch Vorformatierung, Unterstützt USB Hot-Plug und Plug&Play, Ein/Aus-Schalter, LED-Funktionsanzeige, Stromversorgung über mitgeliefertes Netzteil

Was kennzeichnet eine Festplatte?

- Offensichtlich: Kapazität
- Versteckt: IOPs (Ein/Ausgabe Operationen pro Sekunde)
 - Umdrehungsgeschwindigkeit (5,4K – 15K)
 - Zugriffsgeschwindigkeit hängt vom Ort ab, an dem der Block gespeichert ist
 - $IOPs = 1 / (Latenzzeit + Zugriffszeit)$
- Anschlusstechnik (FC, SAS, USB, Firewire, ...)
- Mechanische Qualität, bestimmt durch Nutzungsintensität:
 - Platte dreht nur, wenn nötig (Notebooks)
 - 365 x 24 x 7, 100% Device busy (Enterprise-Niveau)
- SSDs befreien die Speicherung von mechanischen Einflüssen
 - Größenordnungen „schneller“ und „teurer“
 - Wiederbeschreibbarkeit begrenzt

Typische IOPs-Zahlen

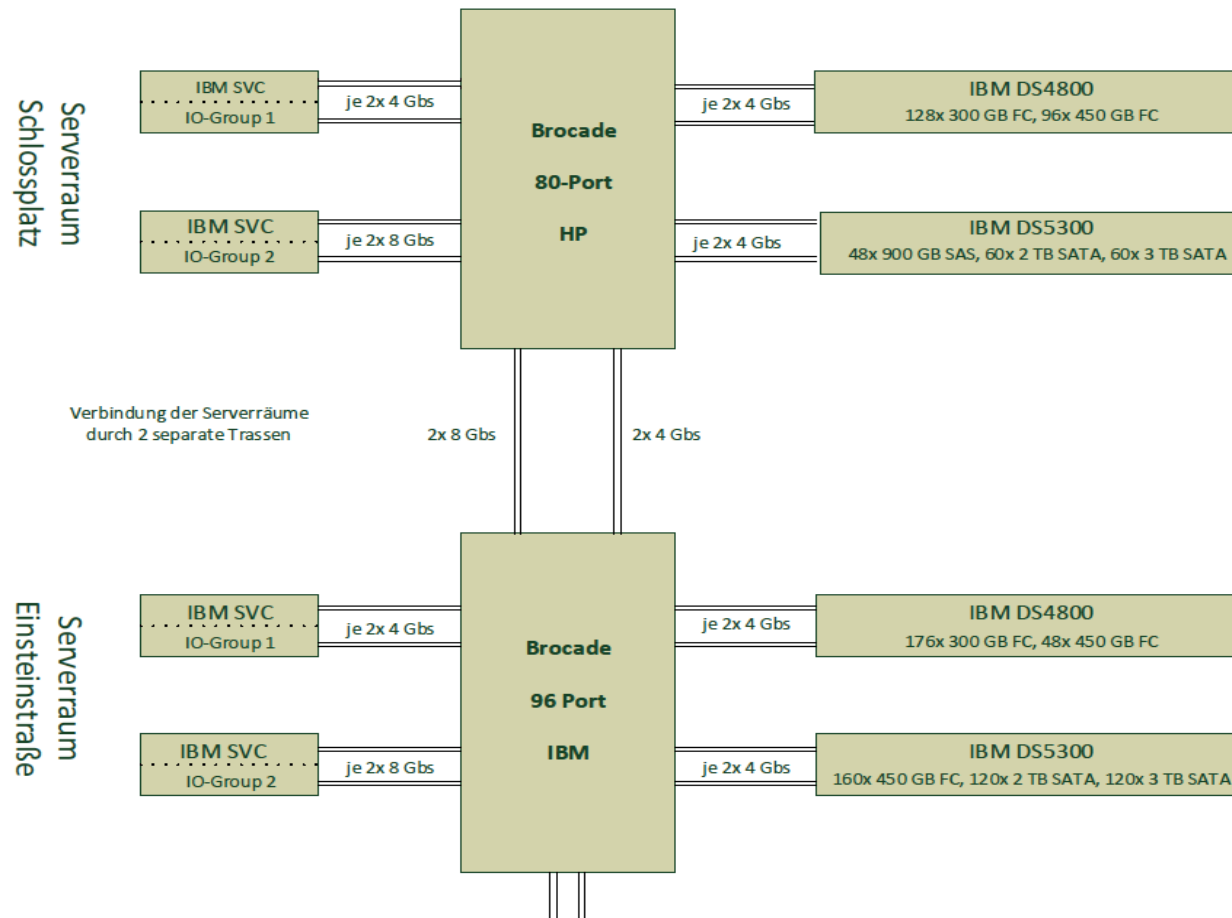
Plattentyp	RPM	IOPs-Bereich
SATA	5,4 K	50-75
SATA	7,2 K	75-100
FC/SAS	10 K	100-125
FC/SAS	15 K	175-200
SSD	N/A	5.000- 10.000

ZIVspeicher – In Zahlen

➤ Stand: November 2012

- 4 Storage-Server
- Laufwerke: 1.112 (ohne PALMA)
 - FC@15K (300 GB, 450 FC)
 - SAS@10K (900 GB)
 - SATA@7,2K (2 TB, 3 TB)
- (Brutto-)Kapazität: 1.228 TB
- IOPs: 156.000
- Einschübe: 47
- Platzbedarf: 5 Racks
- Zu 1/3 genutzt von IVVen
- Genutzt von etwa 500 virtuellen und realen Servern

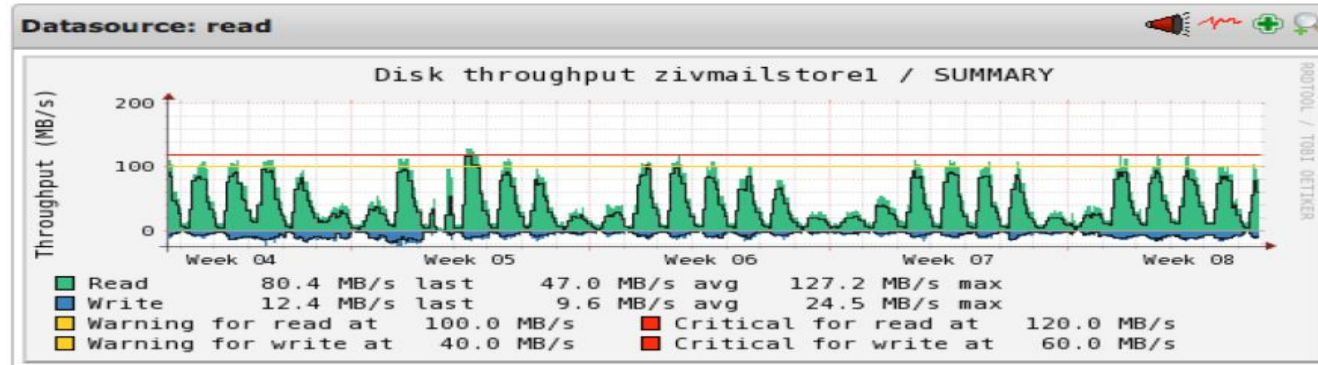
Speicherinfrastruktur ZIV



ZIV-speicher: IO-Bandbreite, Open-Source-Mail, GPFS

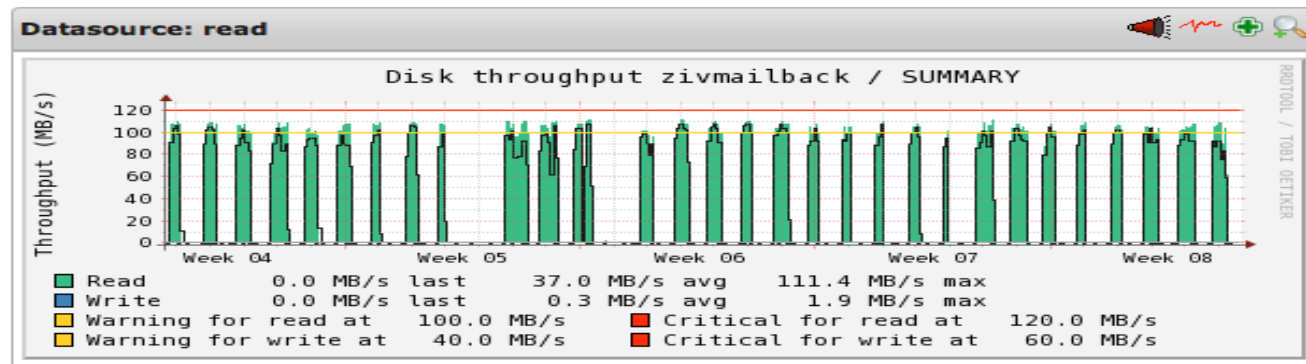
Host: zivmailstore1 **Service:** Disk IO SUMMARY

One Month 21.01.13 14:35 - 22.02.13 14:35



Host: zivmailback **Service:** Disk IO SUMMARY

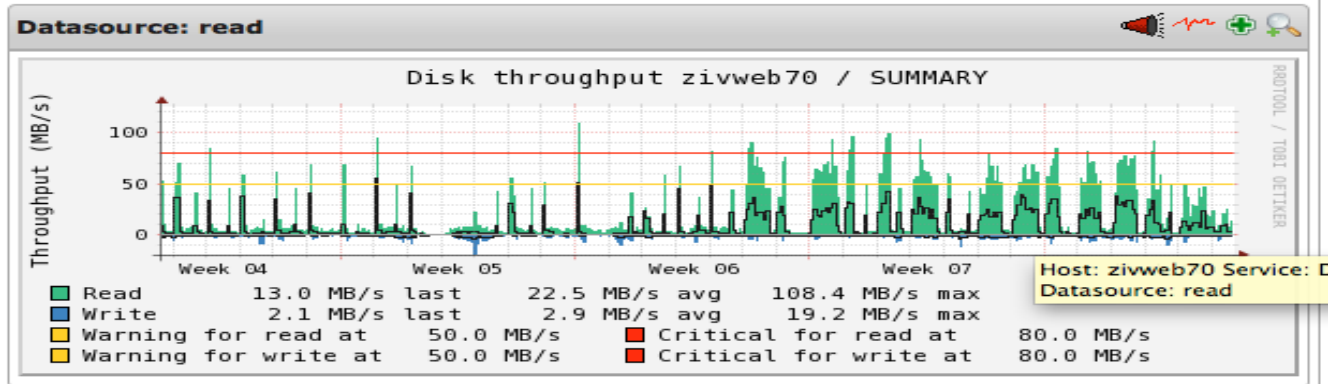
One Month 21.01.13 14:49 - 22.02.13 14:49



ZIVspeicher: IO-Bandbreite, Webserver-Park, GPFS

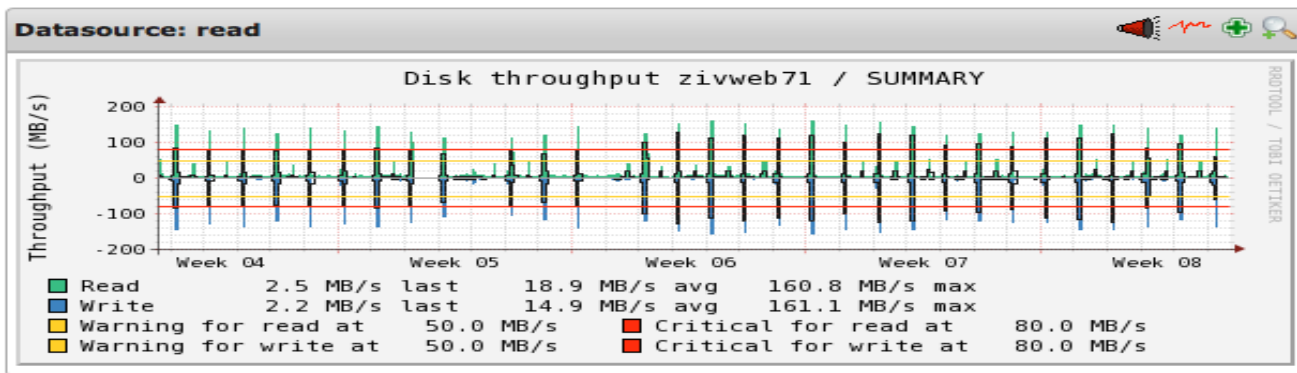
Host: zivweb70 **Service:** Disk IO SUMMARY

One Month 21.01.13 14:44 - 22.02.13 14:44



Host: zivweb71 **Service:** Disk IO SUMMARY

One Month 21.01.13 14:43 - 22.02.13 14:43



Schutz gegen Datenverlust durch defekte Laufwerke

- Spare-Volumes: Ungenutzte Laufwerke, die im Notfall einspringen
- RAID-5: Eine Platte darf ausfallen
 - Ein Spare-Volume übernimmt den Platz des defekten Laufwerks
 - RAID5-Schutz wird neu gebildet (Resync)
 - Nach Austausch des defekten Laufwerks: Kommando „Zurück“
- RAID-6: Zwei Platten dürfen ausfallen
- RAID-Schutz kostet Schreib-Performance
 - 4 IOs bei RAID-5, 6 IOs bei RAID-6
- Optimierungs-Problem
 - Kapazitäts- und IO-Effizienz legen große Volumes nahe
 - Große Volumes haben lange Resync-Zeiten (bis zu Tagen)
- Im ZIV: 1-2 Plattendefekte pro Woche
- RAID-Schutz, der auf ganzen LW beruht, ist nicht mehr ausreichend.

Das Problem der vielen Laufwerke

Technologie & Trends



Forschung, 5 Jahre vor jedermanns Rechenzentrum

Year	System	TFLOPS	GB/s	Nodes	Cores	Storage	Disks
1998	Blue Pacific	3	3	1464	5856	43TB	5040
2000	ASCI White	12	9	512	8192	147TB	8064
2005	ASCI Purple/C	100	122	1536	12.288	2 PB	11.000
2011	HPCS	6000	6000	65.536	512K	120 PB	200.000

Jay Menon, EDGE 2012, Orlando

Mathematische Prognose:

2020	?					125.000 PB	5 mio*
------	---	--	--	--	--	------------	--------

(*) Annahme Disk-Kapazität von 25TB

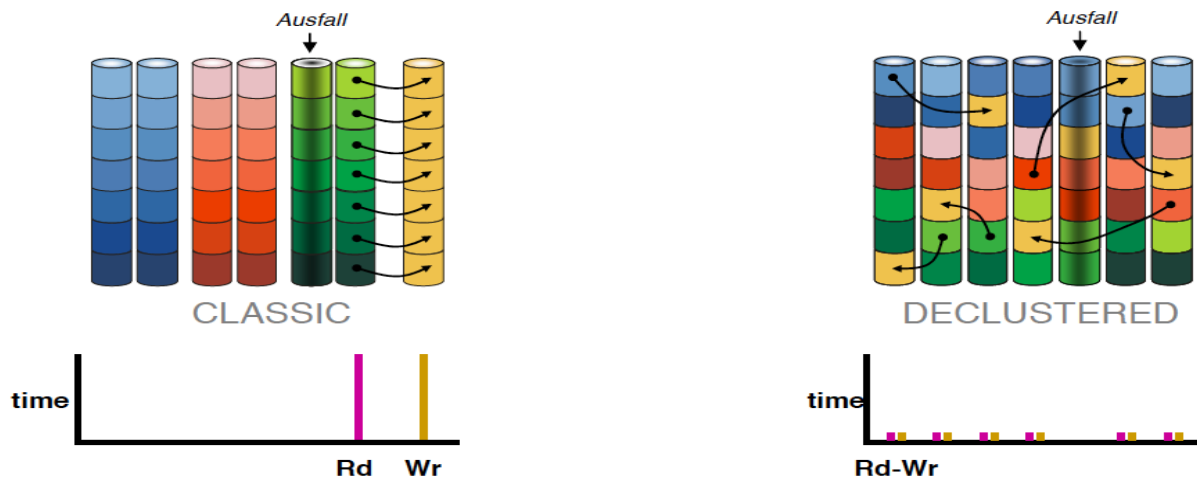
300 Ausfälle pro Tag, MTDDL mit RAID 6 = 2 Wochen

Ein neuer Schutz: Declustered RAID

Technologie & Trends

IBM

Warum multi-TB Platten *Declustered RAID* brauchen



"For 200k drives, we get 12 rebuilds per day, statistically."

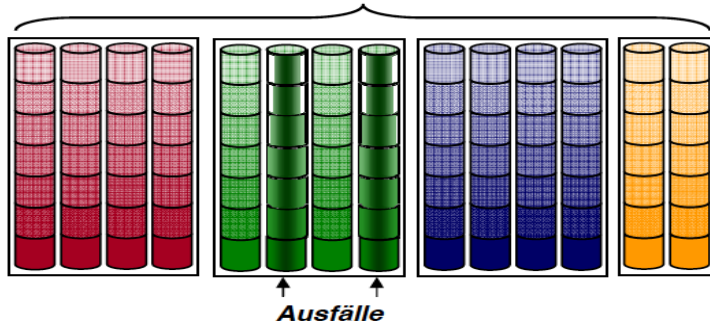
Statistik: Wie lange bis zum Datenverlust?

Technologie & Trends

IBM

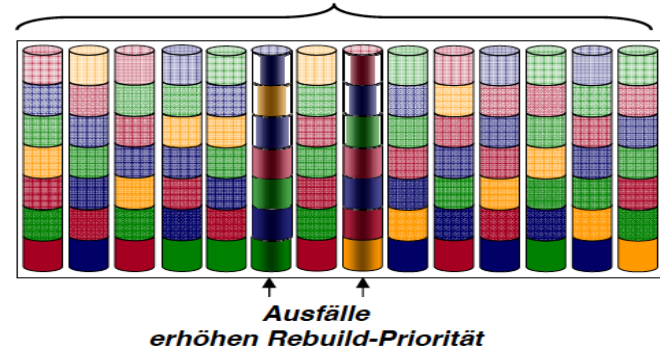
Declassified RAID 6 (RAID-D2)

14 physical disks / 3 traditional RAID6 arrays / 2 spares



VS.

14 physical disks / 1 declassified RAID6 array / 2 spares



RAID-D2 : Mean Time to Data Loss ~ 100 Jahre

RAID-D3 : MTDDL ~ 130 Mio Jahre

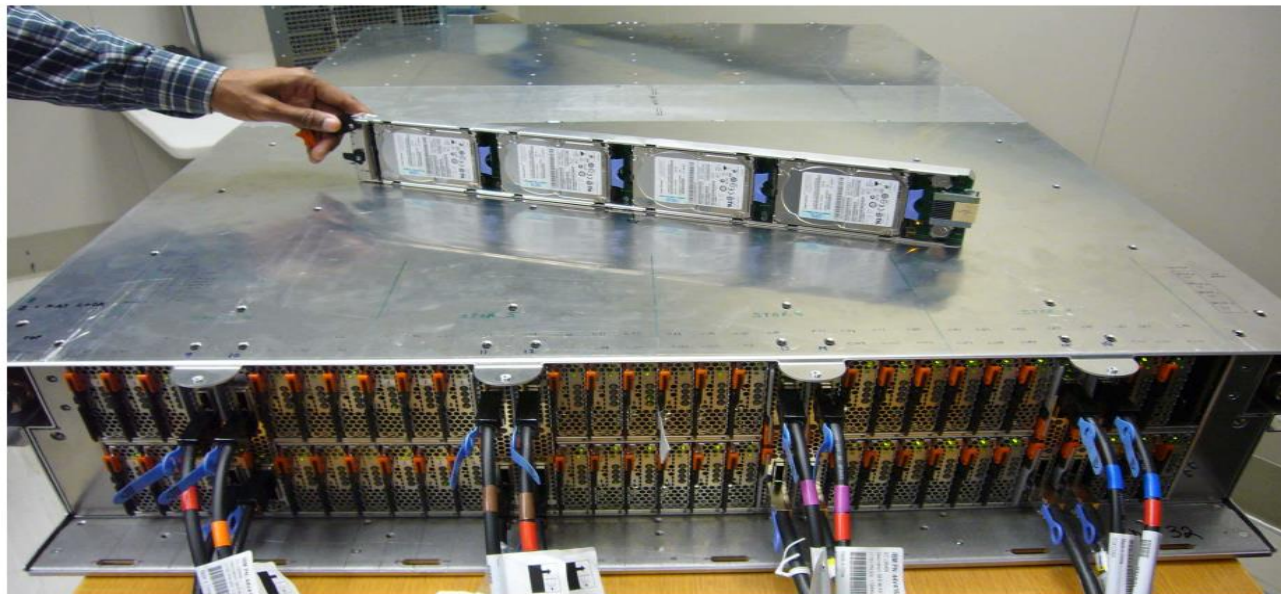
RAID-5 für 100.000 Platten : MTDDL ~ 9 Tage!

Laufwerke in hoher Packungsdichte

Technologie & Trends



120 PB Hardware (hier: ca. 380 TB)



Dense Disk Enclosure – 384 Disks pro Einheit (192 vorn+hinten)

2% Flash SSD für Metadaten

Aktives Laufwerk-Management

- Am Besten: Laufwerke austauschen bevor sie kaputt gehen.
- SMART-Daten der Laufwerke überwachen
- Bei auffälligen Laufwerken:
 - Keine neuen Daten mehr auf die Laufwerke schreiben
 - Lesbare Daten kopieren, nicht-lesbare Daten rekonstruieren
 - Laufwerk als defekt kennzeichnen und Austausch veranlassen
- Speicher-Systeme telefonieren „nach Hause“
 - Ersatzteillieferungen werden automatisch angestoßen
 - Je nach Vertrag: Weiterer Support durch den Hersteller möglich

Unterschiedliche Nutzungsintensität

- Häufige, gelegentliche und seltene Nutzung
 - Häufig → 15K LW, SSDs
 - Gelegentlich → 15K LW während der Nutzungsperiode, 7,2 K LW sonst
 - Selten → 7,2K LW
- Erfahrung: 10% der Daten fordern 90% der Bandbreite
- Servern zugewiesene Datenbereiche (LUNs) enthalten ruhige wie aktive Daten
- Bisher: Eine LUN wird „als Ganzes“ einem Speicherort zugewiesen
 - Gold-Speicher: 15K LW, RAID-5
 - Bronze-Speicher: 7,2 K LW, RAID-6

Selbstoptimierende Speicher

- LUNs werden nicht mehr händisch Gold-Bronze-Kategorien zugeordnet.
- Die Speichersysteme platzieren Daten selbstständig
 - Betrachtet werden Datenblöcke (kleine Teile von LUNs)
 - An Hand ihrer Nutzungshäufigkeit wird der Speicherort festgelegt
- Statistische Analyse braucht Zeit (Tage)
- Ein „eingeschwungenes“ Speichersystem
 - hält aktive Datenblöcke auf schnellen Laufwerken
 - Inaktive Datenblöcke auf langsamen Laufwerken
- Eigene Messungen im Vorfeld der Ausschreibung
 - vSphere-Datenbereichen (VMFS), SVC-Auto-Tiering
 - 2% der Daten machen 95% des IOs

Und was haben wir jetzt beschafft?

➤ Dell-Compellent

- 2x SC8000 Storage Controller
- 192 Laufwerke 2,5“ 300 GB SAS 15K
- 96 Laufwerke 2,5“ 1 TB SAS 7,2K
- 150 TB (Brutto)
- Zusicherung: 45.000 IOPs
- Leistungsaufnahme: 6,2 KW
- Platzbedarf: 28 HE
- SW-Lizenzen: Ab dem 97. Laufwerk keine zusätzlichen Kosten
- 5 Jahre Wartung und Support

➤ Wird im März installiert

Und warum kaufen wir nicht bei Saturn?

- Kapazität ist notwendig aber nicht hinreichend
- IOPs sind wichtig. Wirklich!
- Platzsparende Anschlusstechnik
- Keine 1-zu-1 Zuordnungen zwischen Server, Speicher und Serverraum
- Laufwerke alleine nützen nichts
- Keine Support-Infrastruktur
- Keine zentrale Betriebsüberwachung
- Laufwerke telefonieren nicht „nach Hause“
- Und ... Und ... Und ...