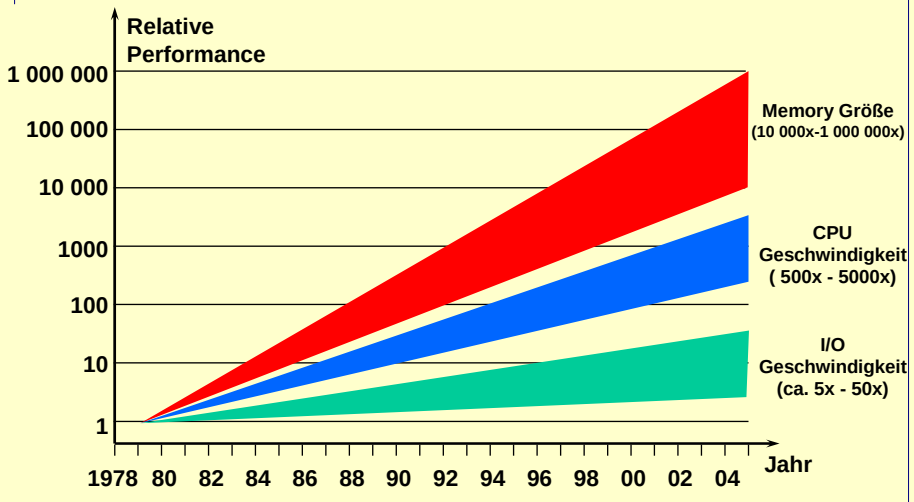
1C02

RAID Performance Grundlagen

Hermann Brunner
Angerwiese 15
85567 Grafing
Te l 080 92 / 328 29
Fax 080 92 / 328 42
hermann@brunner-consulting.de
www.brunner-consulting.de

Wozu überhaupt über I/O Performance nachdenken?



Die wichtigsten Maßzahlen

Deutsch	Englisch	Einheit
Bandbreite	Bandwidth Data Rate / Throughput	**MB/sec**
Datenrate **Durchsatz** I/O Leistung	I/O Request Rate	**I/Os/sec**
Mittlere Suchzeit	Average Seek Time	**msec**
Drehwartezeit Latenzzeit	Rotational Latency	**msec**
Mittlere Zugriffszeit Mittlere Antwortzeit	Average Access Time Average Response Time	**msec**

brunner consulting RAID Performance Grundlagen 3

Bandbreite

(Band Width) [MB/sec]

Beschreibt die Datenmenge, die pro Sekunde bewegt werden kann.

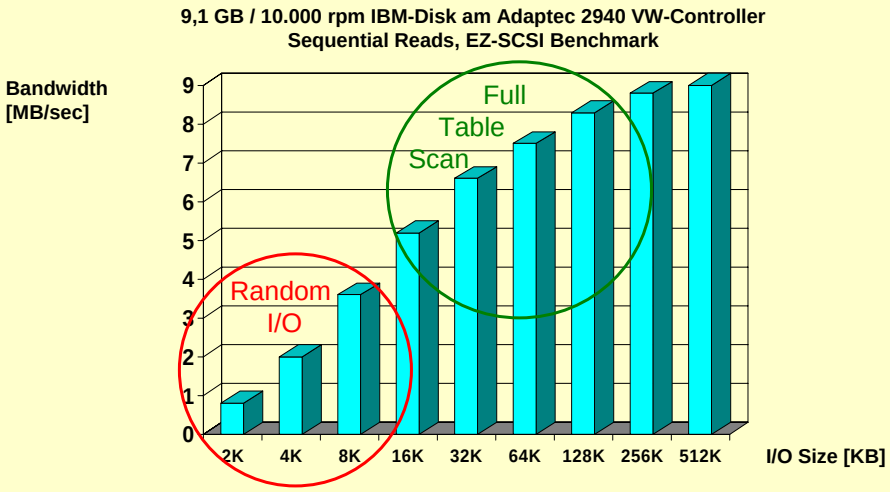
Meßverfahren:

1. Geschwindigkeit, mit der die Leseköpfe die Information von der Oberfläche ablesen (➔ „Peak Data Rate“)
   - ⇨ Nur elektrisch meßbar
   - ⇨ **Kein realistischer Performance-Indikator**

2. Geschwindigkeit, mit der die Daten bei „spiralförmigem Zugriff“ an der Schnittstelle das Laufwerk „verlassen“ (➔ „Spiral Read Rate“)
   - ⇨ Gute Maßzahl für Datenrate
   - ⇨ Wichtiger Performance-Indikator für Anwendungen, die große Datenmengen in riesigen I/Os bewegen, z.B.
     - ❋ BACKUP/RESTORE
     - ❋ Bildverarbeitung
     - ❋ SEISMIK, Meteorologie
     - ❋ Also nur Spezialfälle...

brunner consulting RAID Performance Grundlagen 4

Bandbreite versus I/O-Größe



brunner consulting

RAID Performance Grundlagen 5

Durchsatz

(I/O Request Rate) [IOs/sec]

Beschreibt Anzahl der I/Os pro sec, NICHT die Geschwindigkeit eines I/O

Meßverfahren:

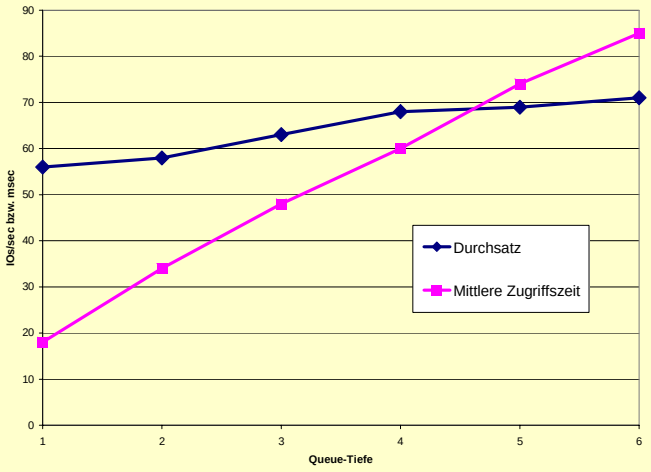
1. Anzahl der I/O-Operationen, die nacheinander abgearbeitet werden können (1 oder mehrere Disks)
2. Anzahl der I/Os bei bestimmten Queue-Tiefen (z.B. 5)
3. Anzahl der I/Os bei mehreren (parallelen) Prozessen
4. Eine Mischung aus 2. und 3.

- Nur Verfahren 1 liefert eine Aussage über die Geschwindigkeit (= Antwortzeit) der einzelnen I/O-Operation
- Verfahren 2, 3 und 4 liefern „bessere“ Zahlen ➔ dadurch beliebter
- Resultate von Verfahren 2, 3 und 4 nur aussagekräftig, wenn gleichzeitig weitere Angaben gemacht werden:
  - ⇨ Mittlere Zugriffszeit
  - ⇨ Queue Tiefe
  - ⇨ Anzahl Dämon-Prozesse

brunner consulting

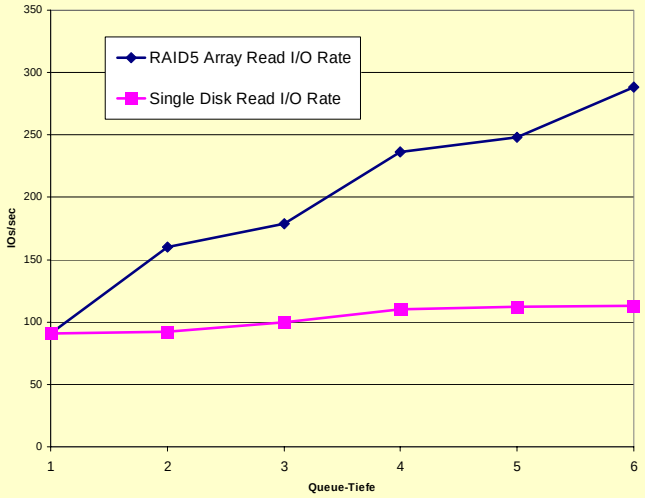
RAID Performance Grundlagen 6

Durchsatz versus Zugriffszeit

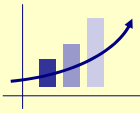


brunner consulting RAID Performance Grundlagen 7

I/O-Durchsatz



brunner consulting RAID Performance Grundlagen 8



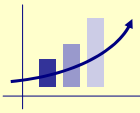
Drehwartezeit

(Rotational Latency) [msec]

Zeit für 1/2 Umdrehung

⇨Eindeutig

⇨Direkt aus der Drehzahl abzuleiten



Mittlere Zugriffszeit

(Average Access / Response Time) [msec]

Wie lange muß ich warten, um meinen I/O zurückzubekommen?

Meßverfahren:

1. Mittlere Suchzeit + Drehwartezeit
2. Wie 1 + Transferzeit für 1 Block

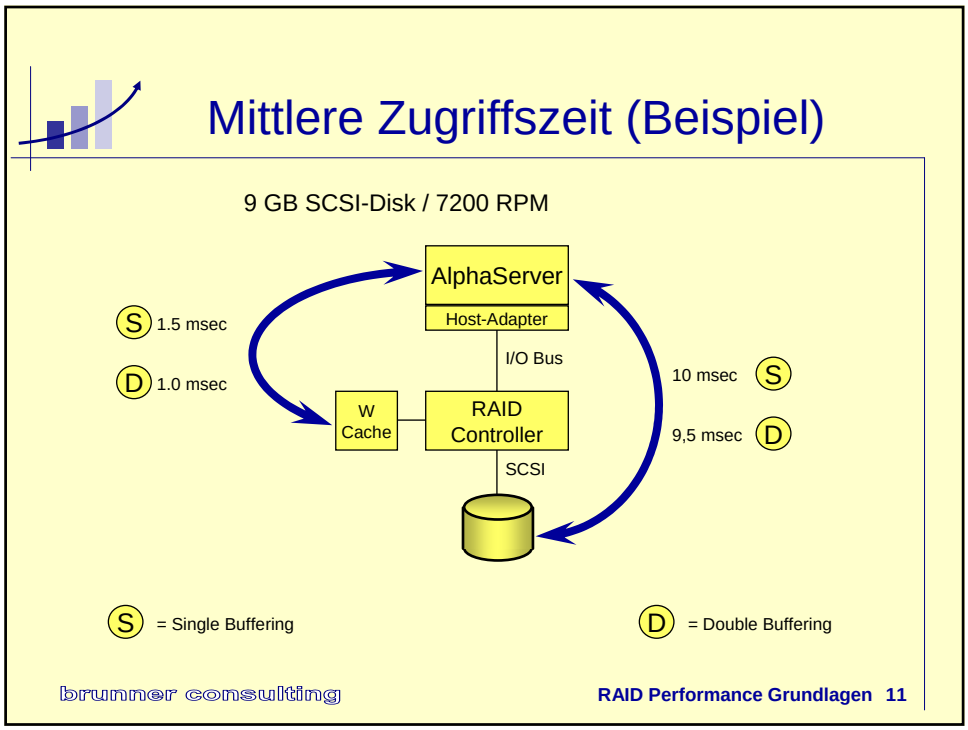
unseriös, aber weit verbreitet

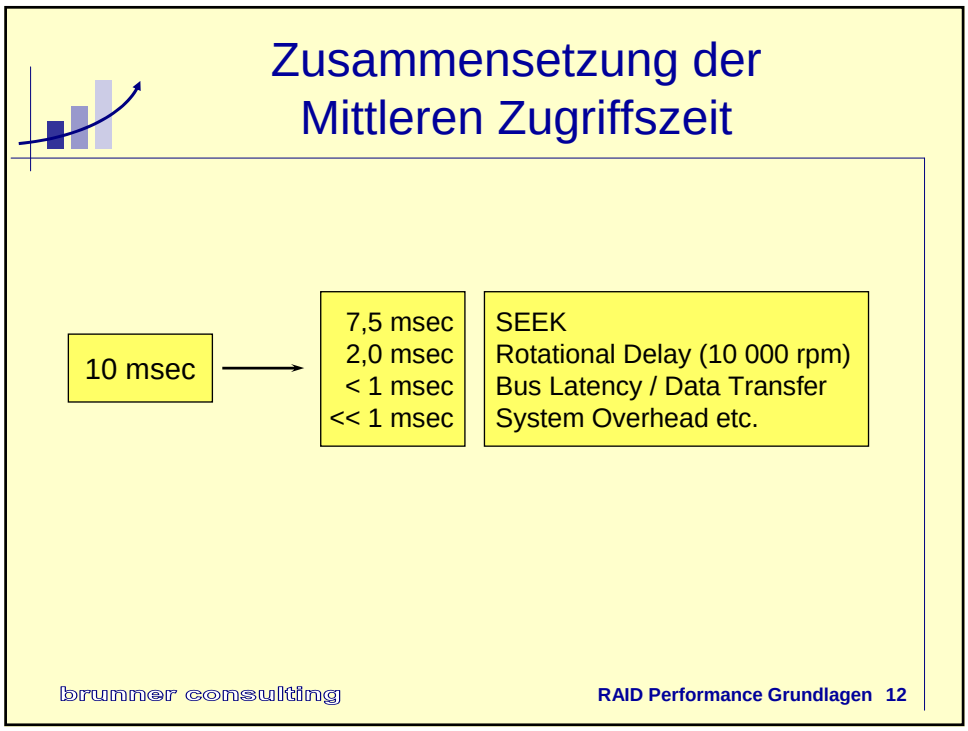
3. Über Benchmark, abhängig von
   - ⇨ Betriebssystem
   - ⇨ Last
   - ⇨ Queueing
   - ⇨ Hardware
   - ⇨ I/O-Größen

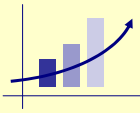
beschreibt I/O-Geschwindigkeit der Gesamtkonfiguration

Probleme:

- ⇨ Ändert sich mit der Last auf verschiedenen Komponenten
- ⇨ Asynchroner I/O


Mittlere Zugriffszeit (Beispiel)
9 GB SCSI-Disk / 7200 RPM
AlphaServer
Host-Adapter
I/O Bus
S 1.5 msec
D 1.0 msec
W Cache
RAID Controller
SCSI
10 msec S
9,5 msec D
S = Single Buffering
D = Double Buffering
brunner consulting
RAID Performance Grundlagen 11

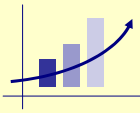

Zusammensetzung der Mittleren Zugriffszeit
10 msec
7,5 msec
2,0 msec
< 1 msec
<< 1 msec
SEEK
Rotational Delay (10 000 rpm)
Bus Latency / Data Transfer
System Overhead etc.
brunner consulting
RAID Performance Grundlagen 12



Achtung:

Die in der Fachpresse verbreiteten I/O Benchmarks prüfen meistens:

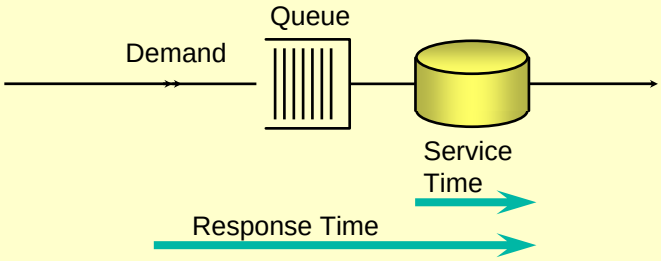
- ⇨ „Maximum Load“
  oder
- ⇨ „Maximum Bandwidth“
  ...aber nur sehr selten
- ⇨ Service Time / Response Time



Dazu Tipps aus der Praxis

- Bandbreiten [MB/sec] sind in interaktiven Environments meist unerheblich.
- Disk Drive Response Zeiten [msec] sind meist der wichtigste Performance-Faktor.
  - ⇨ Ausnahmen: Große LOBs, Full Table Scans, etc.
- Der Geschwindigkeitsverlust von IO-Serving über Netzwerke (NAS!) wird häufig unterschätzt.
  - ⇨ Etwa 50% Erhöhung der Drive Response Time bei einwandfreiem Netzwerk können noch normal sein.
- Drives/Controller mit eingebauten HW-Caches bieten meist **erhebliche** Performance-Vorteile.

Queueing Theorie



Utilization = Service Time ∗ Demand

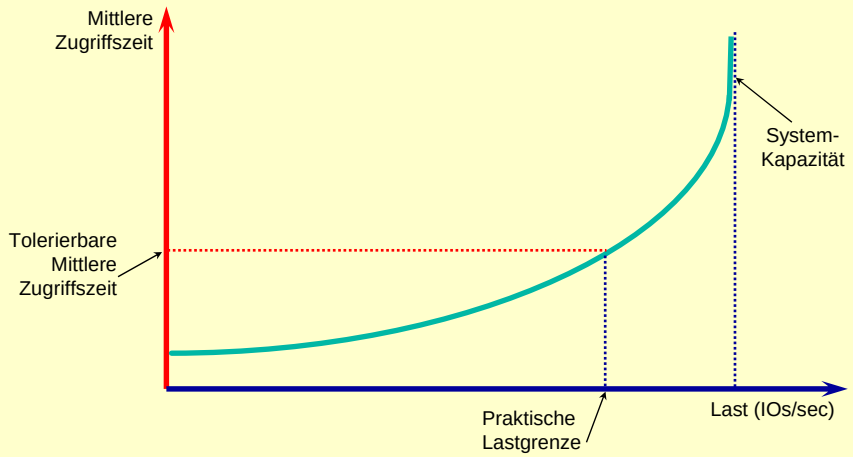
Queue = Utilization / (1 - Utilization)

Response Time = Service Time ∗ (1 + Queue)

Beispiel

Disk	Demand	Service Time	Utilization	Queue	Response Time
RZ26	30 IO/sec	20 ms	60 %	1,5	50 ms
RZ28	30 IO/sec	15 ms	45 %	0,81	27 ms
18GB	50 IO/sec	10 ms	50 %	1,00	20 ms
36GB	50 IO/sec	9 ms	45 %	0,81	16,3 ms
72GB	50 IO/sec	8 ms	40 %	0,67	13,3 ms
Formeln	D	S	U = S ∗ D	Q = U/(1-U)	R = S(1+Q)

I/O-Last versus Zugriffszeit



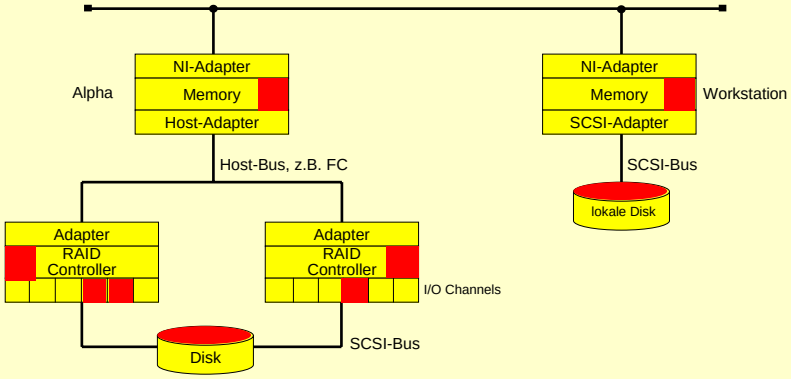
Schlußfolgerungen

Typischerweise wird die Zugriffszeit lange, bevor das Speichersystem die maximale Belastung erreicht hat, inakzeptabel.

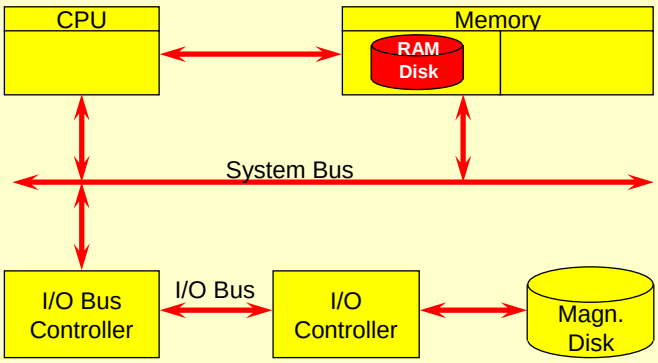
Normalerweise verdoppelt sich die Mittlere Zugriffszeit bereits bei ca. 50% der maximalen Last.

Caching

Einsatz von Caches



Auswirkungen von Caching



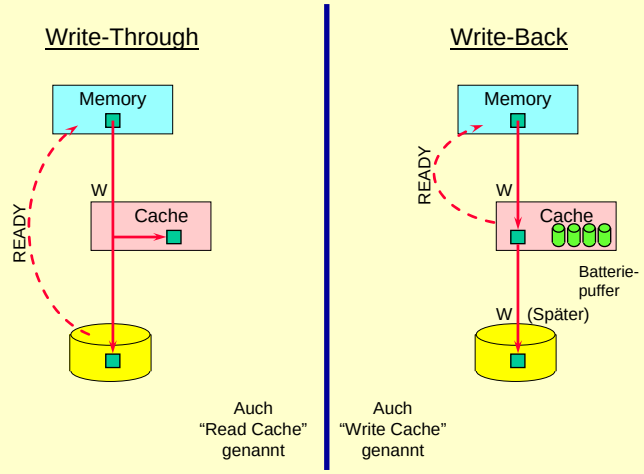
Einsatzmöglichkeiten

- I/O Vermeidung
- I/O Beschleunigung
- Werden wegen fallender Memory-Preise immer interessanter und sind inzwischen (fast) überall im Einsatz
- Bei “Shared Devices” sollte auf der Hardware-Ebene “gecached” werden!
- Nutzen häufig schwer bezifferbar

Performance

- Performance-Gewinne über I/O-Vermeidung
- Meist “Write-Through” Caches, daher Nutzen nur bei READs
- I/O Response Time sinkt um Faktor 10 - 100 (CPU-abhängig)
- I/O-Kanal wird überhaupt nicht benutzt
- Statt dessen Memory ➔ Memory Kopie
  - ⇨ Also **nicht** zu empfehlen bei überlasteter CPU! (Auch nicht, wenn Memory frei ist)

Hardware Caches

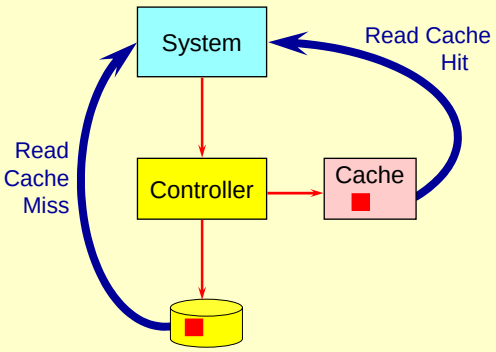


Hardware Caches

- Write Through
  - ⇨ Kein Geschwindigkeitsgewinn beim Schreiben
  - ⇨ Immer konsistente Daten auf der Disk
- Write Back
  - ⇨ Writes in 0,1 – 1,0 msec !!!
  - ⇨ Konsistenz nur über Batterie-Pufferung “einigermaßen” zu gewährleisten!
  - ⇨ Besonders geeignet, um das “Write Performance Hole” bei RAID3 und RAID5 Arrays zu beheben.

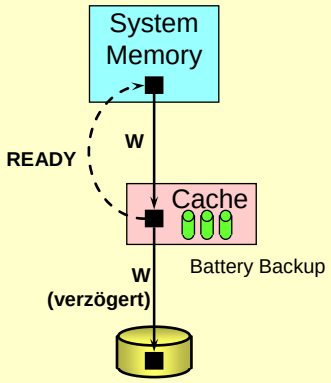
Controller Read Cache

- Vom Host angeforderte Daten werden aus der Cache gelesen (“cache hit”)
- Performance Vorteile
  - ⇨ Schnellere Antwortzeit
  - ⇨ Geringere Device und Device-Bus Utilization
- Cache-Strategie
  - ⇨ “zuletzt benutzt” = LRU optimiert Transaktions-I/Os
  - ⇨ “read-ahead” optimiert große I/Os

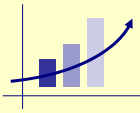


Write Back Cache

- Verhalten wie Read-Cache, plus:
- Host schreibt Daten direkt in die Cache
- Sofortige Rückmeldung an Host
- Daten werden später auf Platte geschrieben
- Performance
  - ⇨ Schnellere Antwortzeit
  - ⇨ Verbessertes RAID WRITE
  - ⇨ Geringere Device und Device-Bus Utilization
- Optimiert Transaktions- und große I/Os
- Daten-Integrität sichern!!!

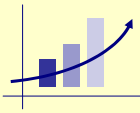


RAID



Motivationen für RAID

- Daten-Sicherheit
- Verfügbarkeit
- Disaster-Toleranz
- Performance
- Management / Handling
- Kosten-Reduktion



Auswahl aus folgenden Optionen

- ⇨Host Based Shadowing
- ⇨Host Based Striping
- ⇨Controller Based Mirroring
- ⇨Controller Based Striping
- ⇨Controller Based Parity RAID

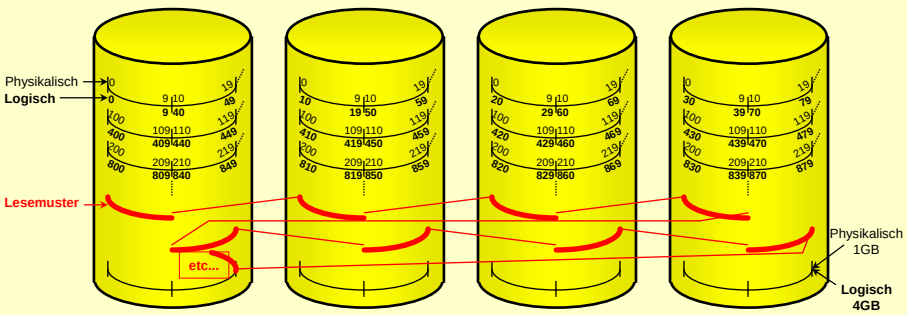
- ⇨Controller Read Caching
- ⇨Controller Write-Back Caching
- ⇨Host Based Caching

Eine Kombination aller Marketing-Folien aller RAID-Anbieter

- ⇨ JBOD okay
- ⇨ RAID 0 gut
- ⇨ RAID 0+1 sehr gut
- ⇨ RAID 3 exzellent
- ⇨ RAID 5 wundervoll
- ⇨ Our RAID 77++ hervorragend!

RAID 0

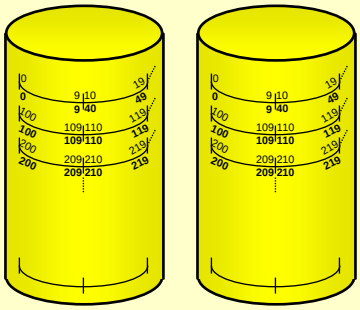
Beispiel: Chunk Size = 10 Sectors
Track Size = 100 Sectors



RAID 1

Shadowing/Mirroring - Daten werden “gespiegelt”

- Vorteile:
  - ⇨ Hohe Sicherheit
  - ⇨ Performance-Gewinn bei Lesezugriffen
  - ⇨ **Wenig** Performance-Verlust beim Schreiben
- Nachteil:
  - ⇨ Teuer

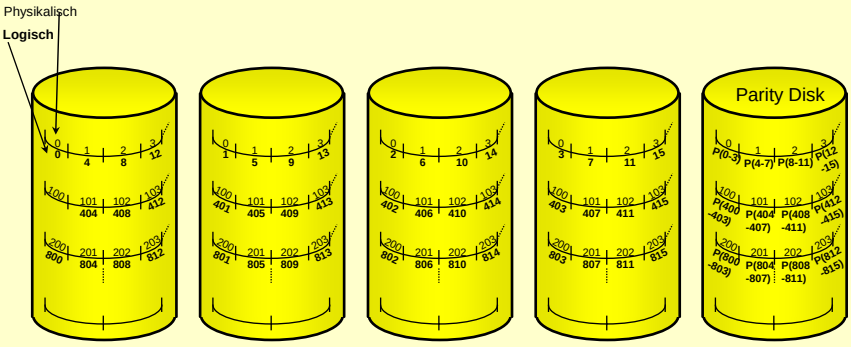


brunner consulting

RAID Performance Grundlagen 33

RAID 3

Beispiel: Chunk Size = 1 Sector
Track Size = 100 Sectors

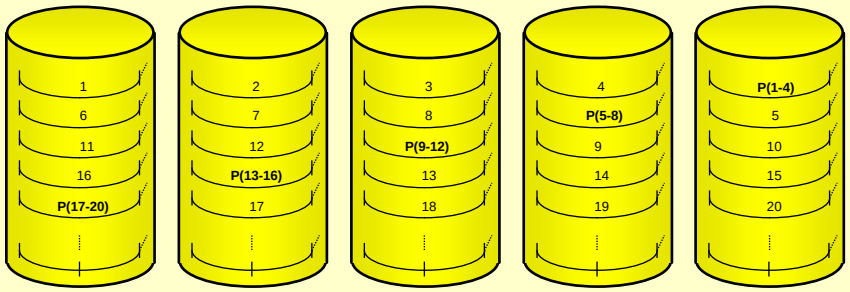


brunner consulting

RAID Performance Grundlagen 34

RAID 5

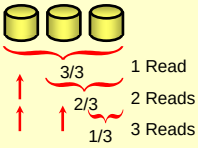
Chunk Size = 1 Track oder größer
Rotierende Parity-Information



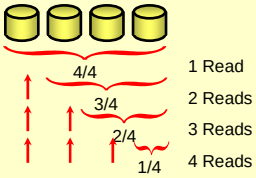
brunner consulting

RAID Performance Grundlagen 35

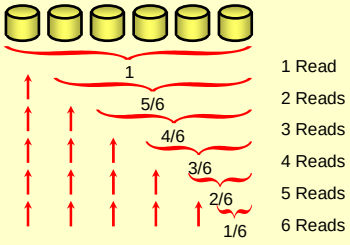
Wahrscheinlichkeit gleichzeitiger Read-Operationen



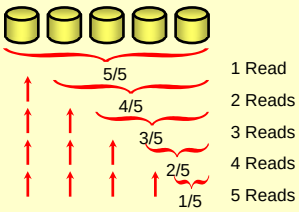
6/3 = 2,0



10/4 = 2,5



21/6 = 3,5



15/5 = 3,0

brunner consulting

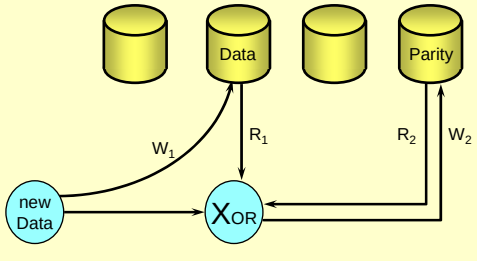
RAID Performance Grundlagen 36

Parallele Reads

Annahme: Array mit 6 Disks

Queue	*Wahrscheinlichkeit paralleler Operationen*	
1	1	1
2	1 + 5/6	1,83
3	1 + 5/6 + 4/6 = 15/6	2,5
4	1 + 5/6 + 4/6 + 3/6 = 18/6	3,0
5	1 + 5/6 + 4/6 + 3/6 + 2/6 = 20/6	3,33
6	1 + 5/6 + 4/6 + 3/6 + 2/6 + 1/6 = 21/6	3,5
∞		6

Parity RAID: Read - Modify - Write



Read:

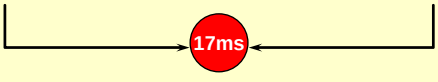
pro Disk: 10 msec

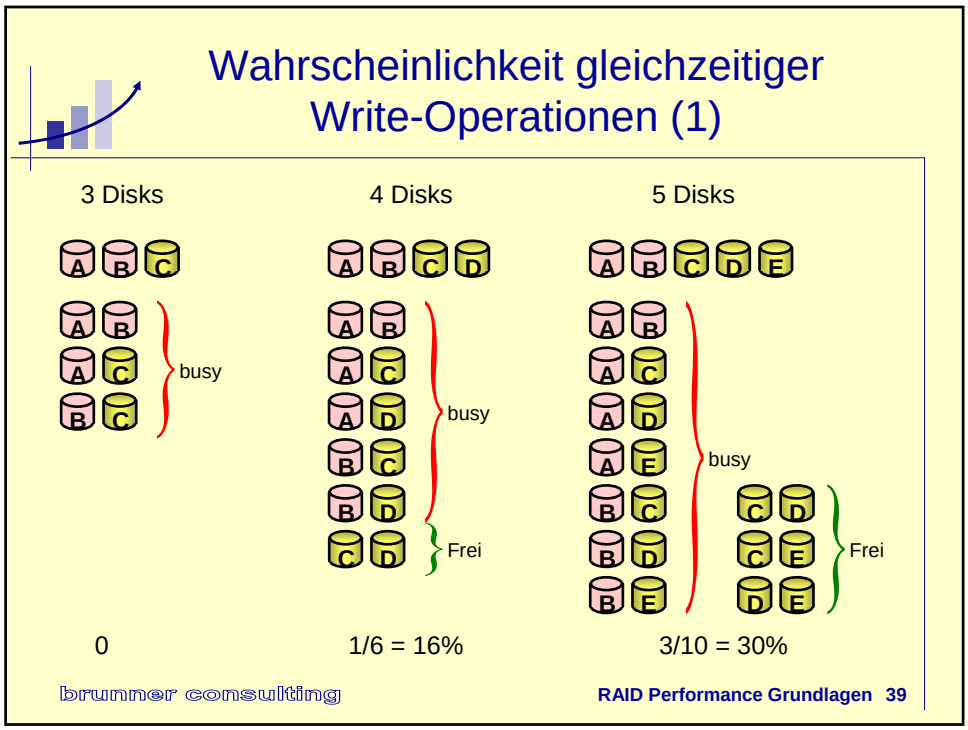
beide: ~ 12 msec

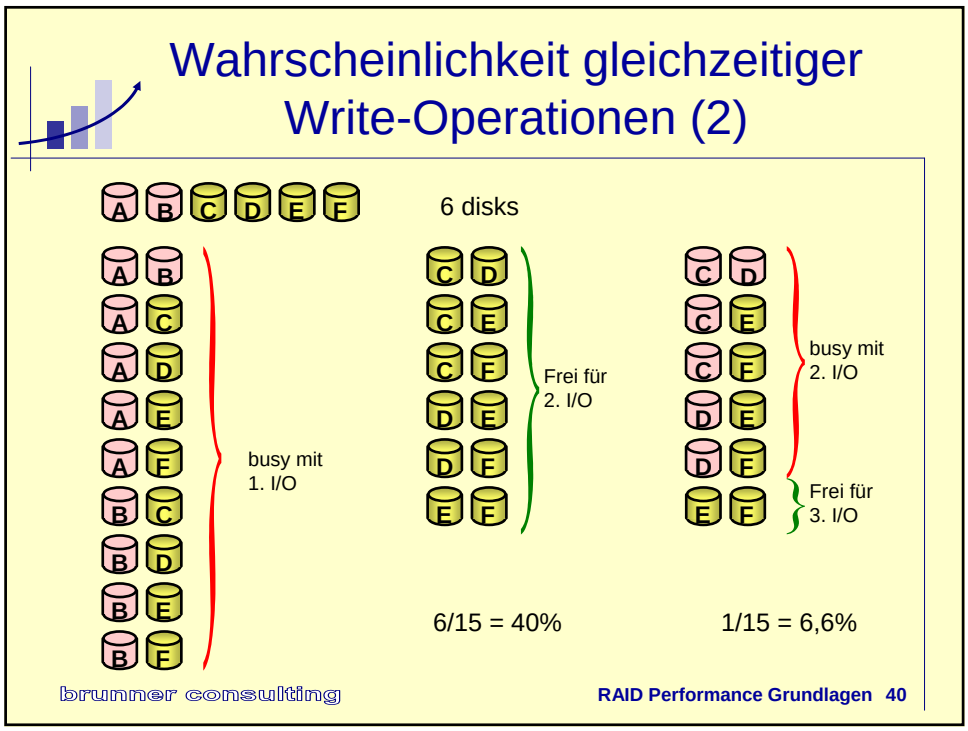
Modify:

"keine Zeit"
(während Warten auf WRITE)

Write: Wartet eine volle Umdrehung
pro Disk 4,33 msec
XFER 0,5 msec
Overhd. 0,1 msec
total: ~ 5 msec




Wahrscheinlichkeit gleichzeitiger Write-Operationen (1)
3 Disks
4 Disks
5 Disks
busy
busy
busy
Frei
Frei
0
1/6 = 16%
3/10 = 30%
brunner consulting
RAID Performance Grundlagen 39


Wahrscheinlichkeit gleichzeitiger Write-Operationen (2)
6 disks
busy mit 1. I/O
Frei für 2. I/O
busy mit 2. I/O
Frei für 3. I/O
6/15 = 40%
1/15 = 6,6%
brunner consulting
RAID Performance Grundlagen 40

Wahrscheinlichkeit gleichzeitiger Operationen

Disks im Array	Gleichzeitige R-M-W Zyklen	Gleichzeitige Reads
2	1,0	1,5
3	1,0	2,0
4	1,166	2,5
5	1,30	3,0
6	1,46	3,5

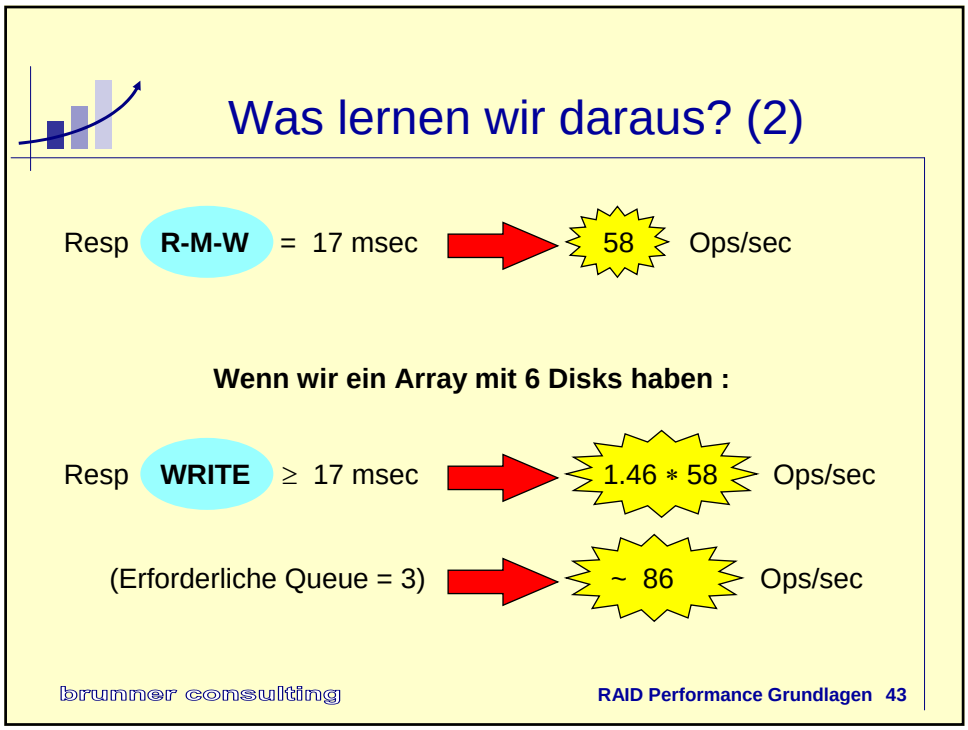
Was lernen wir daraus? (1)

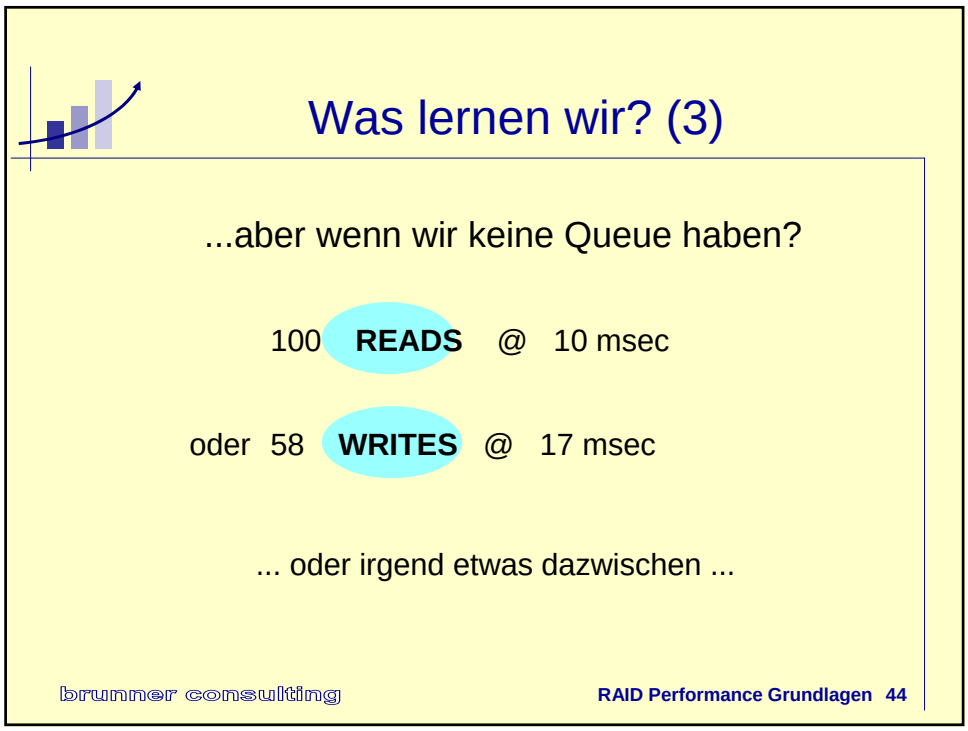
Resp **READ** = 10 msec → 100 Ops/sec

Wenn wir ein Array mit 6 Disks haben:

Resp **READ** ≥ 10 msec → 3,5 * 100 Ops/sec

(Erforderliche Queue = 6) → ~ 350 Ops/sec


Was lernen wir daraus? (2)
Resp
R-M-W
= 17 msec
58
Ops/sec
Wenn wir ein Array mit 6 Disks haben :
Resp
WRITE
≥ 17 msec
1.46 * 58
Ops/sec
(Erforderliche Queue = 3)
~ 86
Ops/sec
brunner consulting
RAID Performance Grundlagen 43


Was lernen wir? (3)
...aber wenn wir keine Queue haben?
100
READS
@ 10 msec
oder 58
WRITES
@ 17 msec
... oder irgend etwas dazwischen ...
brunner consulting
RAID Performance Grundlagen 44

Performance der RAID Levels

	Transaktions-I/O		Large File I/O	
	Read	Write	Read	Write
JBOD	OK	OK	OK	OK
RAID 0 (Strip)	Sehr gut	Sehr gut	Gut	Gut
RAID 1 (Shad)	Gut	OK	OK	OK
RAID 0+1	Exzellent	Sehr gut	Gut	OK
RAID 3	Schlecht	Schlecht	Sehr gut	Sehr gut
RAID 4	Sehr gut	Sehr Schlecht	Sehr gut	OK
RAID 5	Sehr gut	Schlecht	OK	OK
RAID 6	Exzellent	Sehr Schlecht	Sehr gut	OK

Vorteile dieser Tabelle?

- ⇨ Sie beschreibt Stärken und Schwächen der RAID Level
- ⇨ Sie unterscheidet zwischen

  ✵ Transaktions-I/O
  (klein, häufig, zufällig, Queues, Antwortzeit-orientiert) und

  ✵ Großen Files / Stream I/O
  (große, sequentielle, Bandbreiten-orientierte)

Nachteile der Tabelle?

- Sie beschreibt theoretisches Verhalten des RAID Mapping Algorithmus
- Sie ignoriert CACHES

aber ... heute...

- alle RAID-Anbieter benutzen WRITE CACHES, um die “schlechte” und “sehr schlechte” Performance des PARITY RAID bei Transaktions-Writes zu umgehen

Impact of Read and Write Caches

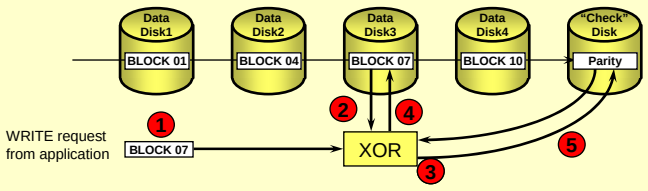
	XACTION I/O (Small)		Large File I/O	
	Read	Write	Read	Write
JBOD	OK	Sehr gut	OK	Gut
RAID 0 (Strip)	Sehr gut	**Exzellent**	Gut	Sehr gut
RAID 1 (Shad)	Gut	Sehr gut	OK	Gut
RAID 0+1	**Exzellent**	**Exzellent**	Gut	Sehr gut
RAID 3	**Schlecht**	Gut	Sehr gut	**Exzellent**
RAID 4	Sehr gut	Gut	Sehr gut	Gut
RAID 5	Sehr gut	Sehr gut	OK	Gut
RAID 6	**Exzellent**	Sehr gut	Sehr gut	Gut

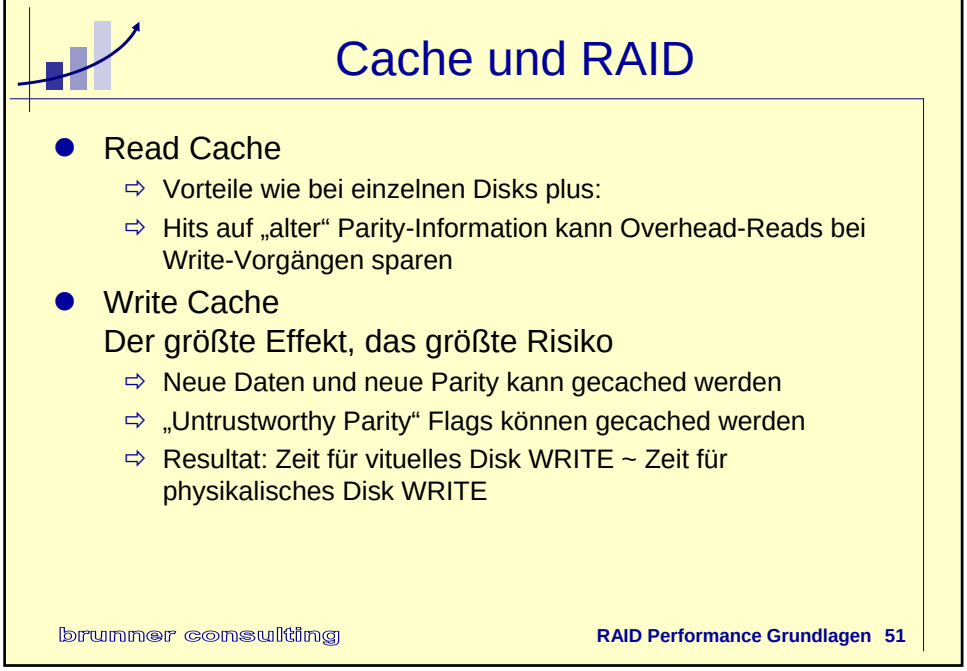
Wichtige Faktoren der RAID Performance

- **Parallelität** in der Anwendung
  - ⇨ Mehrere parallele I/Os?
  - ⇨ Synchrone/Asynchrone I/Os?
- **Read/Write-Rate** der Anwendung
- **Read-Update-Zyklen** in Anwendung helfen Parity-RAID-Controllern, die READ-MODIFY-WRITE Operationen zu optimieren
- **I/O-"Muster"**
  - ⇨ Hat die Read-Cache eine ernsthafte Chance, hohe HIT-Raten zu erreichen?
- **Cache**-Implementation

Write Cache und RAID

- Schreiben zu einem RAID-System
  1. Daten vom Host holen
  2. Lesen vom alten BLOCK 07 und Daten prüfen
  3. Berechnen vom alten BLOCK 07 XOR alte Parity-Daten XOR neuer BLOCK 07
  4. Schreiben vom neuen BLOCK 07
  5. Schreiben der neuen Parity-Daten
- Schritte 2, 3, 4, and 5 können der „Fertig"-Meldung folgen

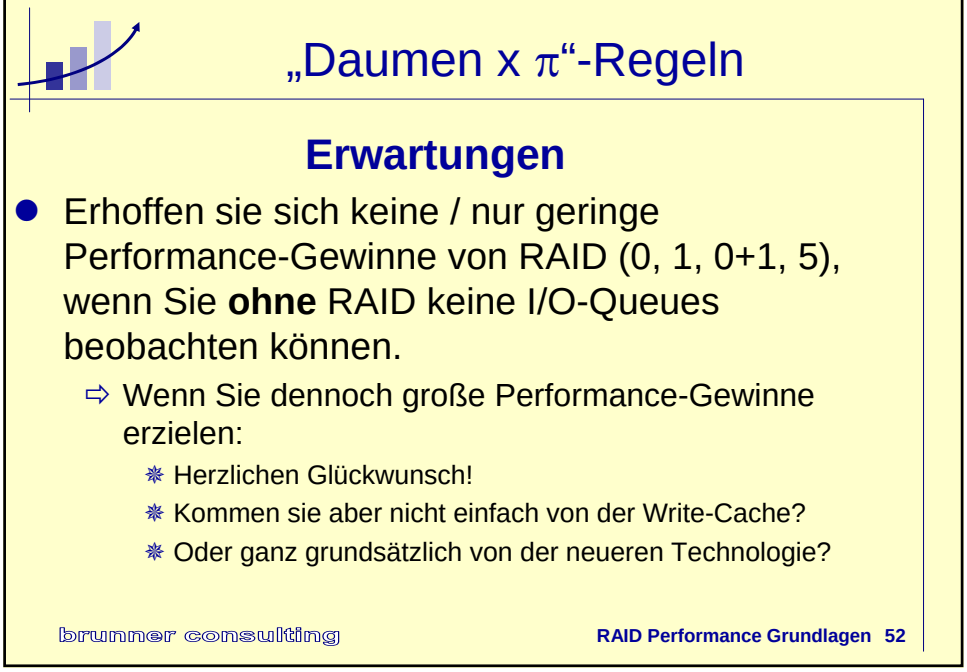




Cache und RAID

- Read Cache
  - ⇨ Vorteile wie bei einzelnen Disks plus:
  - ⇨ Hits auf „alter“ Parity-Information kann Overhead-Reads bei Write-Vorgängen sparen
- Write Cache
  Der größte Effekt, das größte Risiko
  - ⇨ Neue Daten und neue Parity kann gecached werden
  - ⇨ „Untrustworthy Parity“ Flags können gecached werden
  - ⇨ Resultat: Zeit für vituelles Disk WRITE ~ Zeit für physikalisches Disk WRITE

brunner consulting RAID Performance Grundlagen 51



„Daumen x π“-Regeln

Erwartungen

- Erhoffen sie sich keine / nur geringe Performance-Gewinne von RAID (0, 1, 0+1, 5), wenn Sie **ohne** RAID keine I/O-Queues beobachten können.
  - ⇨ Wenn Sie dennoch große Performance-Gewinne erzielen:
    - ❋ Herzlichen Glückwunsch!
    - ❋ Kommen sie aber nicht einfach von der Write-Cache?
    - ❋ Oder ganz grundsätzlich von der neueren Technologie?

brunner consulting RAID Performance Grundlagen 52

Planen Sie genau und gründlich!

- Entscheidungen für ein bestimmtes RAID-Verfahren sind häufig “unumkehrbarer” als Sie vorab denken würden...
  - ⇨ Backup - Restore von riesigen Arrays macht man nicht so gerne / oft.
  - ⇨ Der berühmte **“Wo geht’s hier zum Bahnhof bitte?”** -Effekt schlägt erbarmungslos zu.
  - ⇨ Erklären Sie mal Ihrem Chef, daß Sie jetzt noch die Host-Based Shadowing Lizenzen nachkaufen müssen, weil es die RAID5-Option nicht gebracht hat.

Welches RAID?