

Re: Measuring IOPS and Raid penalty

Source:

<http://www.tech-archive.net/Archive/Exchange/microsoft.public.exchange.admin/2006-03/msg04439.html>

- *From:* SW <SW@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx>
 - *Date:* Fri, 31 Mar 2006 02:58:02 -0800
-

Just an update for you. We did the move last weekend and now have 3 RAID 20 volumes on 15k ROM disks and split the database, the transaction logs are all on their own spindles. Before we had a single database and it was on a RAID 5 config. Before our Physical disk avg writes per second for 0.037 (37ms), 0.042 (42ms) just to random days and MOM 2005 used to email us all the time about 50ms latency on DB's. Now over the disks we get 0.025, 0.019, 0.022 and MOM 2005 doesn't email me anymore, well I had one email :)

Wondered what you thought?

"John Fullbright" wrote:

Why 3 storage groups and 6 databases? Is that a firm requirement? The current MS guidance for Exchange 2003 SG layout is to start with four storage groups even if each SG only contains one database (KB 890699). The reason for this is that log buffers are allocated on a per storage group basis, and by having more storage groups and thereby log buffers, you reduce log contention. I think your environment is small enough that log contention won't be a major issue. Each storage group has a separate set of logs, and you do want to place each set of logs on a separate set of physical spindles. You need 8 spindles in a RAID 10 set to support your databases. If you use the vss provider and snapshots as part of your backup, the choice of one big drive or several smaller ones will impact the granularity of backup/restore. With the 14 spindles I would:

- RAID 10 4 drives SG1 Databases
- RAID 10 4 drives SG2 Databases
- RAID 1 2 drives SG1 Logs
- RAID 1 2 drives SG2 Logs
- RAID 1 2 drives SMTP queue and badmail directories
- Boot volume and page file – local disk

By using a LUN for each storage group, your granularity of backup/restore will be at the SG level if you use snapshots. You could have one or more databases inside a storage group. It's tempting to go with just one, but remember that eseutil requires 110% of the space of your largest database available if you ever need to do a repair or offline defrag. By splitting

Re: Measuring IOPS and Raid penalty

into several smaller databases within the storage group, you reduce the required amount of space to reserve for Eseutil operations. For example:

If I have one 45GB Store in SG1, the size LUN I would need in order to be able to run eseutil would be 99GB. If I split the 45GB store into four 11.25GB stores, then the size of the LUN drops to 57.375GB. When you run a repair or defrag, eseutil creates a temporary db. At the end of the process, it deletes the current db and renames the temp db. Yes you could use the "t" parameter to map to a share somewhere or another drive, by if you do this the rename operation becomes a copy. This will lengthen the repair or defrag process by the amount of time it takes to copy the database over the wire; something you want to avoid if possible.

"SW" <SW@xxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in message
news:0EE34511-2A71-483A-9A8E-633D8F197B11@xxxxxxxxxxxxxxxxxxxx

John, we have a bank of 14 disks. 6 are taken up by our RAID 5 volume (10,000 RPM) leaving us 8 slots. We have bought 14 15,000 disks. We have a 90GB private database. Have 2 solutions please advise:

Solution 1:

We have 750 mailboxes, but have 250 disabled users, so it will be 500 users.

3 RAID 10 volumes and split the database:

- 4 x Disks (one storage group with 2 databases)
- 4 x Disks (one storage group with 2 databases)
- 4 x Disks (one storage group with 2 databases)

First we would create 2 RAID 10 volumes and split the database over to these
then remove the RAID 5 volume to create the last RAID 10 volume.

or

Solution 2:

1 RAID 10 volume and split the database:

- 8 x Disks (3 storage groups with 6 databases)

We are not sure if have more RAID 10 volumes with many databases is better
than just one big RAID 10 volume with many databases.

Thanks in advance

Re: Measuring IOPS and Raid penalty

"John Fullbright" wrote:

How many users? What kind of users? What does your environment look like?

I've been in many environments, and my experience is that the MS assumptions for light, medium and heavy users are low. I always measure. My experience has been that the average IOPS/user is closer to 1.5. Things that impact that number:

Mapi applications (Blackberry, Goodlink, Unity, Rightfax, etc.)
Desktop search, email client search applications (Google, Lookout, etc.)
Average mailbox size
Average number of items per folder
Attachment size limits

I have seen many environments where there are no limits, voicemail and fax are integrated with exchange, all users have a crackberry, and the standard desktop deploy image contains a desktop search engine. In these environments, the measured IOPS/user has been in excess of 3. Clearly the assumption of 1 IOP/heavy user is nowhere close. You combine that with RAID 5 and an IOPS/spindle assumption which does not account for the IO response time, and you have an underperforming mess.

Clearly, RAID 10 will give you better performance than RAID 5 for the same type and number of spindles. Moving to 15K RAID 10, with a 657 requirement

Re: Measuring IOPS and Raid penalty

and a 2:1 read/write ratio means you will need enough spindles to reach 867.14 or 8 spindles in your RAID 10 set. To reach the measured 657 requirement with RAID 5 and a 2:1 read/write ratio you would need 13 spindles in your RAID 5 set. That's a 38% reduction in spindle count using RAID 10 vs. RAID 5.

"SW" <SW@xxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in message news:0A622870-6EE9-4DBD-8CB0-7D4564426B7A@xxxxxxxxxxxxxxxxxxxx

Also is 657 high?

"SW" wrote:

John, based on my info I have given you is this current IOPS too high. It's a RAID 5 with 6 10,000 ROM disks. We will go to a RAID 10 with 15,000 disks in hope of seeing an improvement, what do you think?

"John Fullbright" wrote:

$$750 * .73 = 547.5$$

Add 20%

Re: Measuring IOPS and Raid penalty

$547.5 * 1.2 =$
657 IOPS
for
Database

Now
calculate
the logs.
Next,
calculate
your
read/write
ratio and
apply the
write
penalty for
your RAID
type.

For
Example:

Database
disk
requirement
with a
measured
read/write
ratio of 2:1.

Reads
 $= 657 * .66 =$
433.62
Writes =
 $657 * .33 =$
216.81

The write
penalty for
RAID 10 is
2:

Required
IOPS
capacity =
 $433.62 + (216.81 * 2)$
 $= 867.14$

That's about
10 spindles

Re: Measuring IOPS and Raid penalty

at 85
IOPS/spindle
or 8
spindles at
120
IOPS/spindle

The write
penalty for
RAID 5 is
4:

Required
IOPS
capacity =
 $433.61 + (216.81 * 4)$
= 1300.86

For raid 5,
first we
figure out
the number
of data
spindles
 $-1300.86/85$
=
15.3 for
10K SCSI
or
 $1300.86/120$
= 10.845 for
15K SCSI.
Then you
need
to
add in
parity
spindles.
1/6 on HP
or up to
1/11 on
EMC.

RAID 5 on
HP with
10K
spindles =
19
RAID 10 on
HP with
10K

Re: Measuring IOPS and Raid penalty

spindles =
10
RAID 5 on
HP with
15K
spindles =
13
RAID 10 on
HP with
15K
spindles = 8
RAID 5 on
EMC with
10K
spindles 18
RAID 10 on
EMC with
10K
spindles =
10
RAID 5 on
EMC with
15 K
spindles =
13
RAID 10 on
EMC with
15K
spindles = 8
spindles

Of course
for NetApp
ther is no
write
penalty:

RAID 4
10K = 9
spindles
RAID 4
15K = 7
spindles
RAID DP
10K = 10
spindles
RAID DP
15K = 8
spindles

John

Re: Measuring IOPS and Raid penalty

=

"SW"

<SW@xxxxxxxxxxxxxxxxxxxxxxxxxxxx>

wrote in

message

news:A3D7EDF6-E941-4E8C-AA14-CE0E0A26E5FC@xxxxxxxx

If

I

use

our

peak

then

that

is

0.730:

IOPS/mailbox

=

(average

disk

transfer/sec)

÷

(number

of

mailboxes)

our

average

disk

transfer/sec

peak

was

0.730

and

the

number

of

mailboxes

are

750

Would

this

Re: Measuring IOPS and Raid penalty

equal
730/750
=
097
or
0.730/750
=
9.7

"John
Fullbright"
wrote:

"IOPS/mailbox
=
(average
disk
transfer/sec)
÷
(number
of
mailboxes)"

According
to
"Optimizing
Storage
Performance
for
Exchange
Server
2003"
you
should
be
using
peak,
not
average.
Think
about
it.
If
you
use
the
average,
you'll

Re: Measuring IOPS and Raid penalty

be
undersized
and
performing
poorly
50%
of
the
time.

If
my
minmum
is
1,
my
average
is
5,
and
my
peak
is
10,
then
if

I
design
for
average

...
This
is
a
common
sizing
mistake.

From
the
paper

"8.
Identify
the
ex2003base
server
that
experienced
the
highest
load.

Re: Measuring IOPS and Raid penalty

Use the data collected from the server with the highest load as your server/processor/storage baseline.

Use the following best practices:

.
Always design your system to allow 20 percent more utilization than you expect for peaks. This allows the storage and processors to handle spikes during peak periods.

.

Re: Measuring IOPS and Raid penalty

Megacycles
per
mailbox
and
IOPS
per
mailbox
change
as
the
server
configuration
changes.
The
following
list
includes
potential
factors
that
can
change
the
given
megacycles
per
mailbox
and
IOPS
per
mailbox.

.
Mailbox
sizes
are
changed
significantly

.
Max
message
size
is
changed
significantly

.
Third
party
applications

Re: Measuring IOPS and Raid penalty

are
added
or
removed

.
Exchange
features
are
added
or
removed.

.
Average
concurrency
of
the
users
changes
(more
or
less
users
are
online
using
the
system
at
any
given
time).

9.
After
the
spreadsheet
(included
with
the
download
of
the
guide
Optimizing
Storage
for
Exchange
Server

Re: Measuring IOPS and Raid penalty

2003)
is
populated
and
your
mailbox
profiles
are
determined,
you
can
design
your
storage
solution.
For
example,
if
your
analysis
indicates
that
your
standard
mailbox
profile
translates
to
..75
IOPS
per
mailbox
and
1.25
megacycles
per
mailbox,
you
can
determine
the
following
requirements
for
a
4,000
mailbox
server:

.
Mailbox

Re: Measuring IOPS and Raid penalty

Count:

4,000

.

Peak

DB

IOPS:

(4,000

×

.70)

=

3,000