abstract

In this talk we will examine how the OpenEdge RDBMS uses disk storage and the many configuration choices.

Some choices are better than others and we discuss the pros and cons of fixed and variable extents, extent sizes, the difference between logical and physical volumes, partitions, logical volume managers, SSD, RAID 10, RAID 5 and 6, SAN, NAS, iSCSI, and many other exciting and rewarding topics.

PUG CHALLENGE AMERICAS EXCHANGE

Guide To Database Storage

Gus Björklund, Ronin

PROGRESS BravePoint

We are

Notices

- Please ask questions as we go

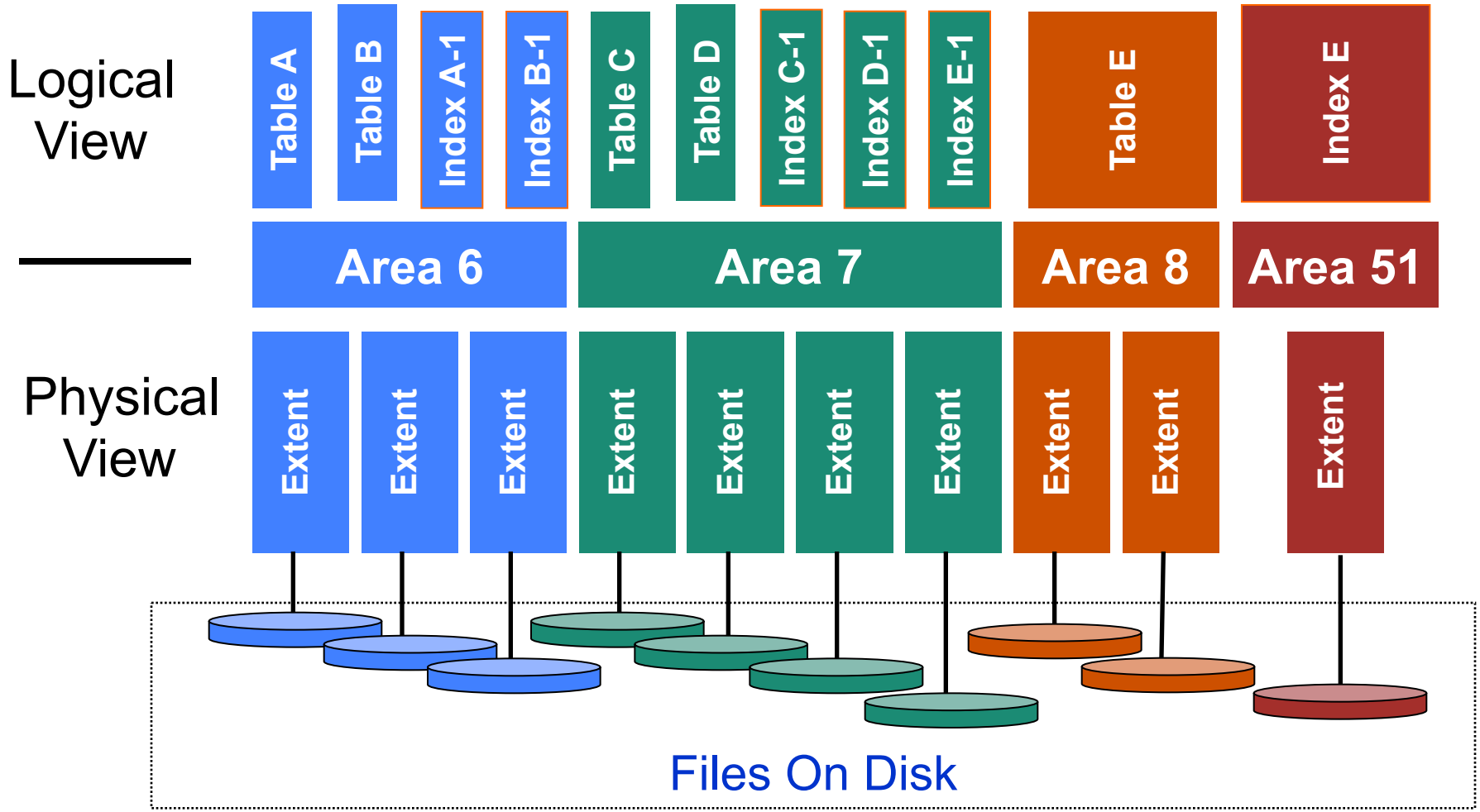

HD
INDEX

list of things

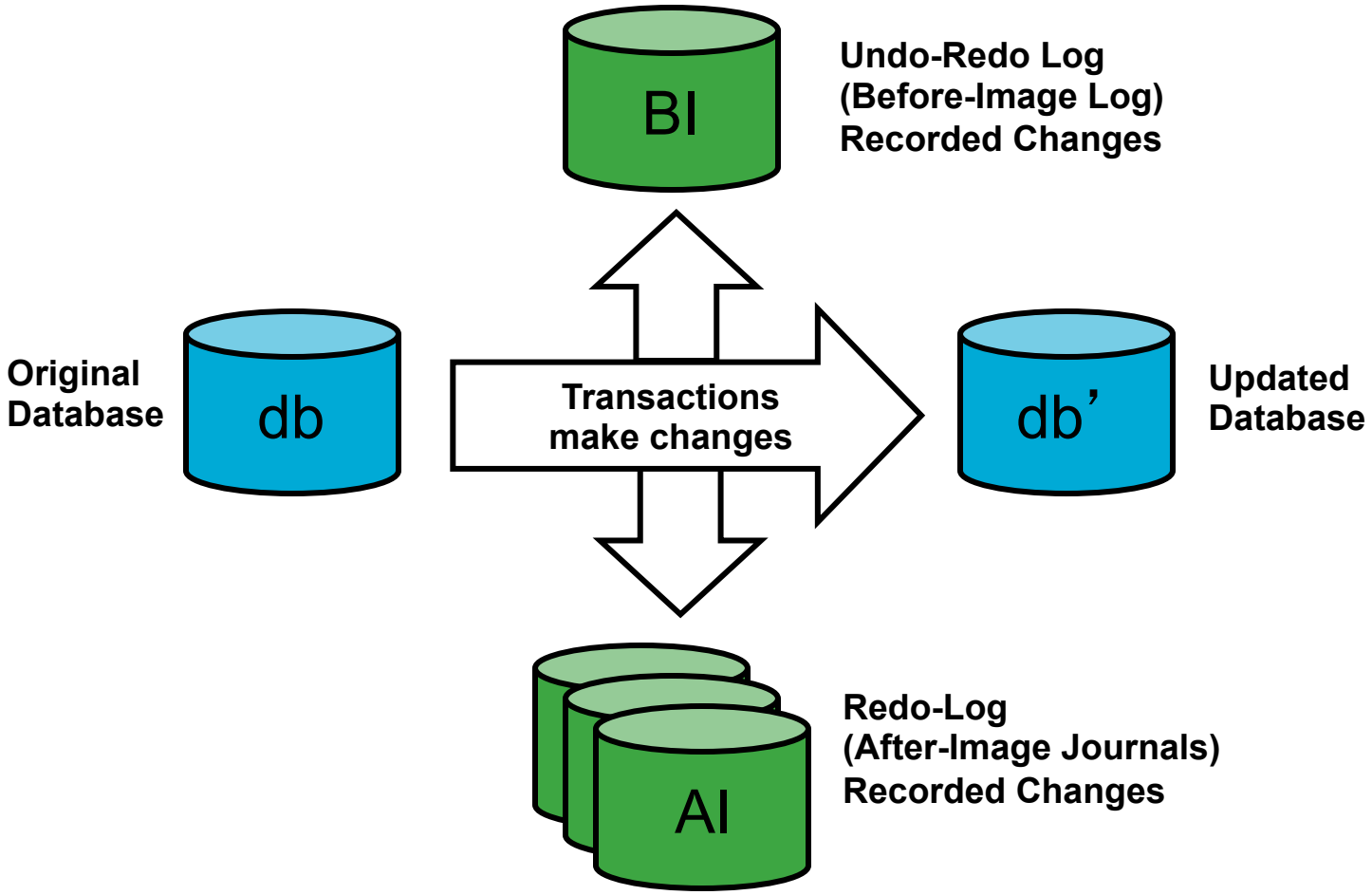
- what has to be stored ?
- storage requirements
- storage device types
- using multiple devices
- device configurations
- OpenEdge RDBMS configuration
- not to do list
- todo list

what do we need to store ?

Progress RDBMS Storage Areas



Progress RDBMS Transaction Logs



Database Storage Requirements

Database Storage Requirements

- Reliability
- Reliability
- Reliability

Database Storage Requirements

- Reliability
- Reliability
- Reliability
- Performance
- Flexibility
- Cost Effectiveness

Progress RDBMS expects

- pread (), pwrite (), fsync (), fdatasync () and other I/O related system calls work properly
- Writes are not buffered by device
- Data block writes are atomic
- OS, Drivers, storage subsystems do not corrupt data

storage device types

Rust geometry

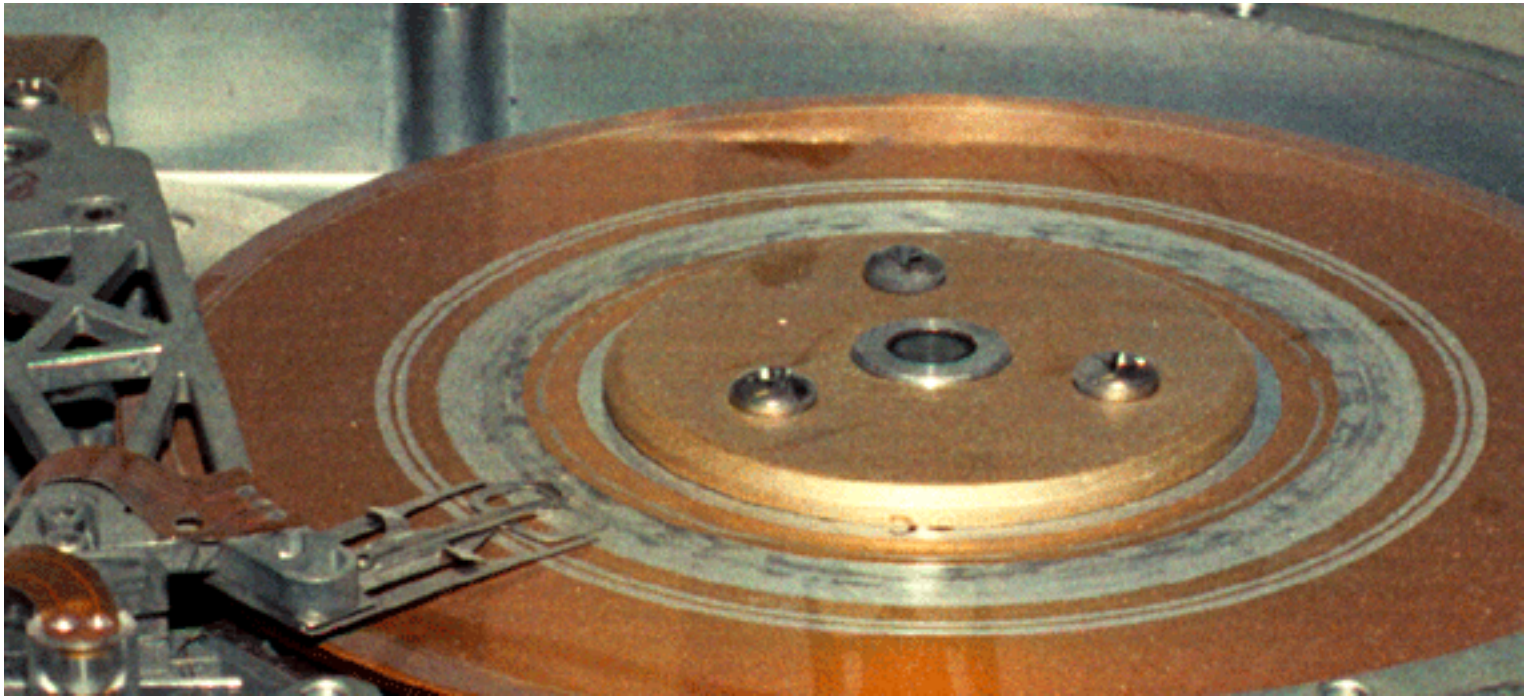


photo: computer museum, university of amsterdam

Speed

- I/O time controlled by basic physics
  - total write time = memory -> controller transfer time + seek time + rotation time + controller -> disk transfer time
- Typical seek times are 1 to 12 milliseconds
  - depends on how far you seek
- Transfer times
  - Memory -> controller
    - almost negligible
    - > 160 megabytes per second
- Controller -> disk
  - depends on disk rpm

Access Times Vary

- Seek + Rotation + Transfer
- Tracks not equal
  - track length is 2πr
  - angular velocity constant
  - linear velocity 2x higher at edge
  - transfer more bits per unit time

How Fast Are Disks ?

	4200 rpm	7200 rpm	10000 rpm	15000 rpm
rev / sec	70	120	166	250
avg rot delay	7 ms	4 ms	3 ms	2 ms
avg seek + avg rot	13 ms	10 ms	9 ms	8 ms
min seek + avg rot	8 ms	5 ms	4 ms	3 ms

Typical “Server Grade” Disks

- Spin faster: 15,000 rpm
- Lower density: 450,000 bpi
- Long duty cycle: 24 x 7
- Higher MTBF: ?
- Faster seeks: 3.5 ms (avg)
- Cost more: $200 (and up)
- Interface: SAS

SSD

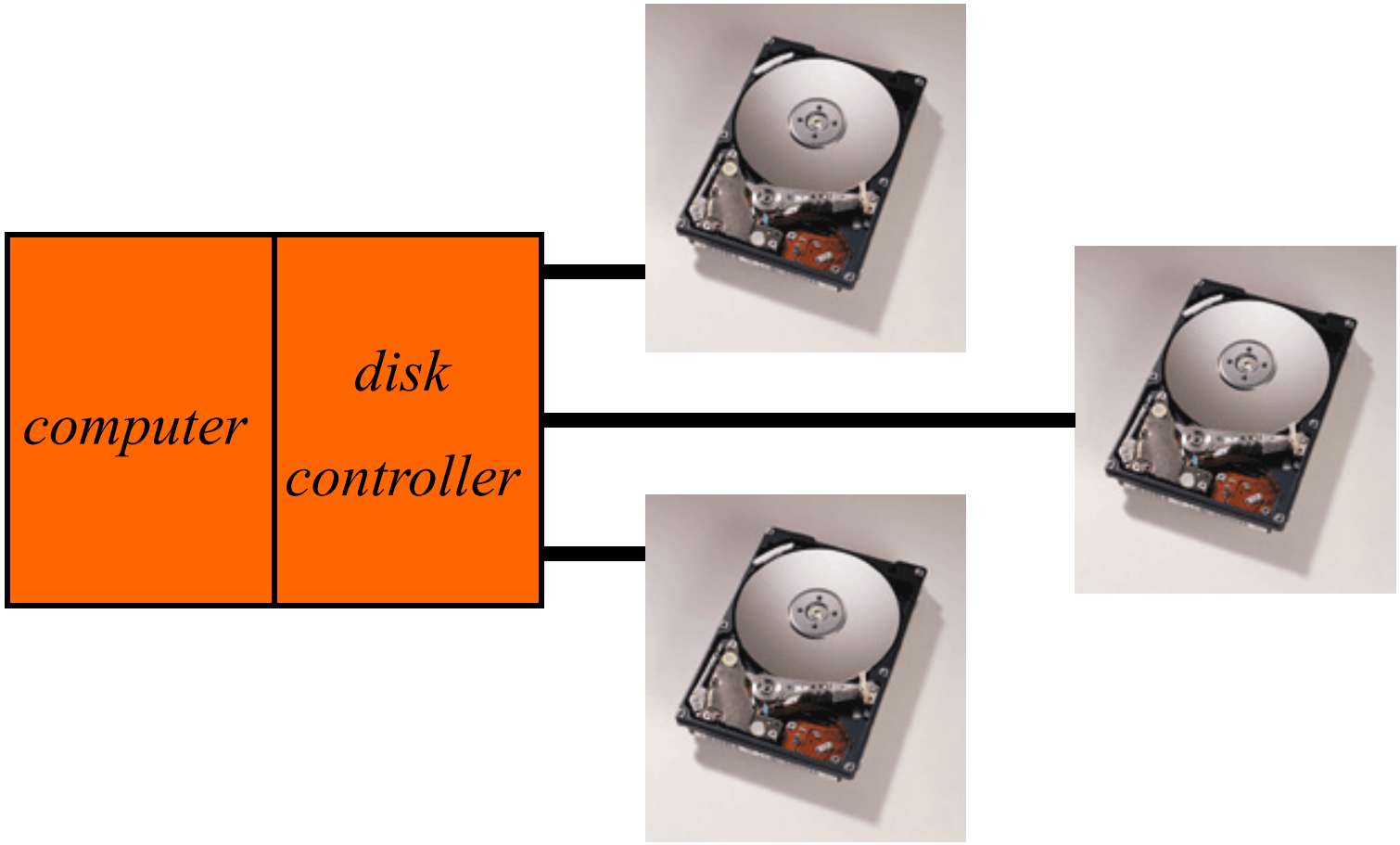
- Faster
- No moving parts
- reliable

using multiple devices

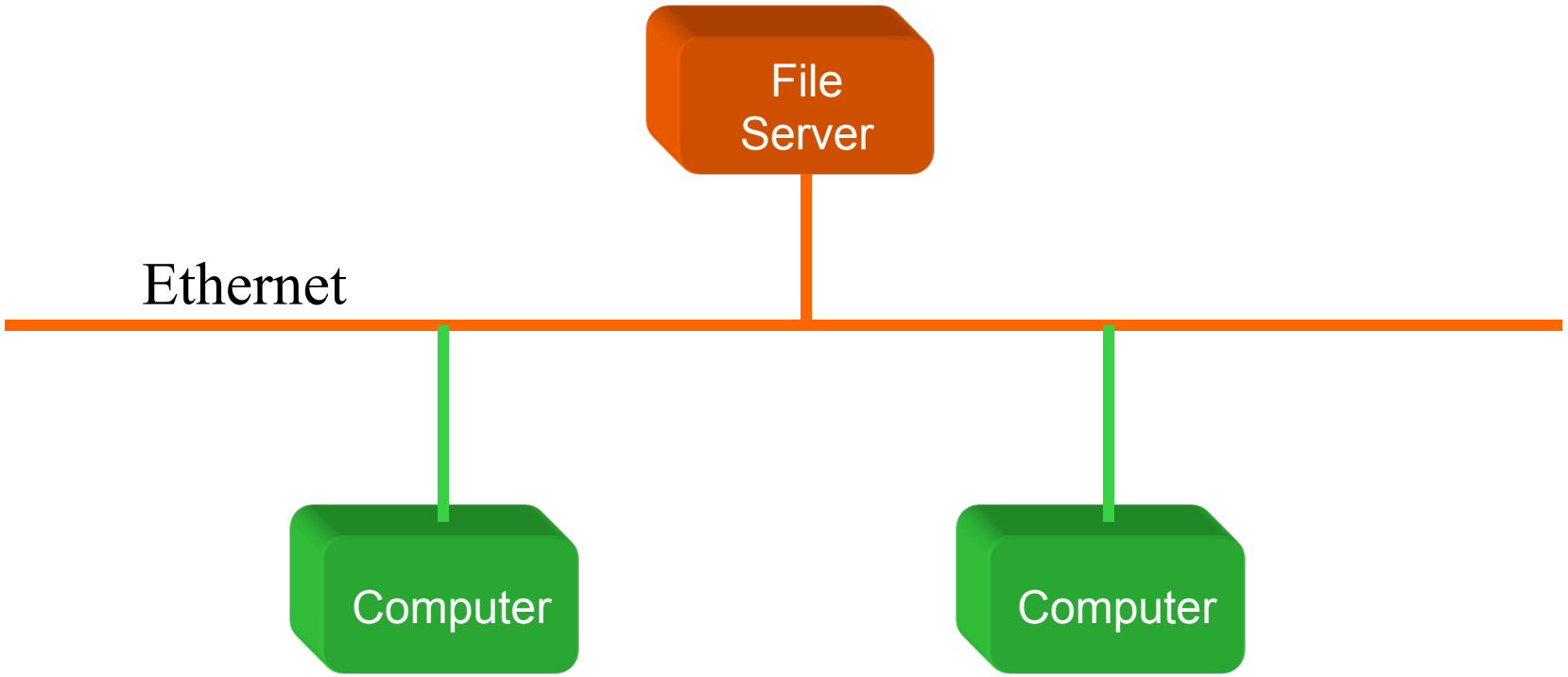
Multiple disks

- One disk can do one transfer at a time
  - Big or small, one at a time
- Two disks can do two transfers at a time
- Controller can handle 4 fast drives at full load
  - 40 MB/sec x 4 = 160 MB/sec
  - More drives at less than full load

JBOD - Just A Bunch Of Disks

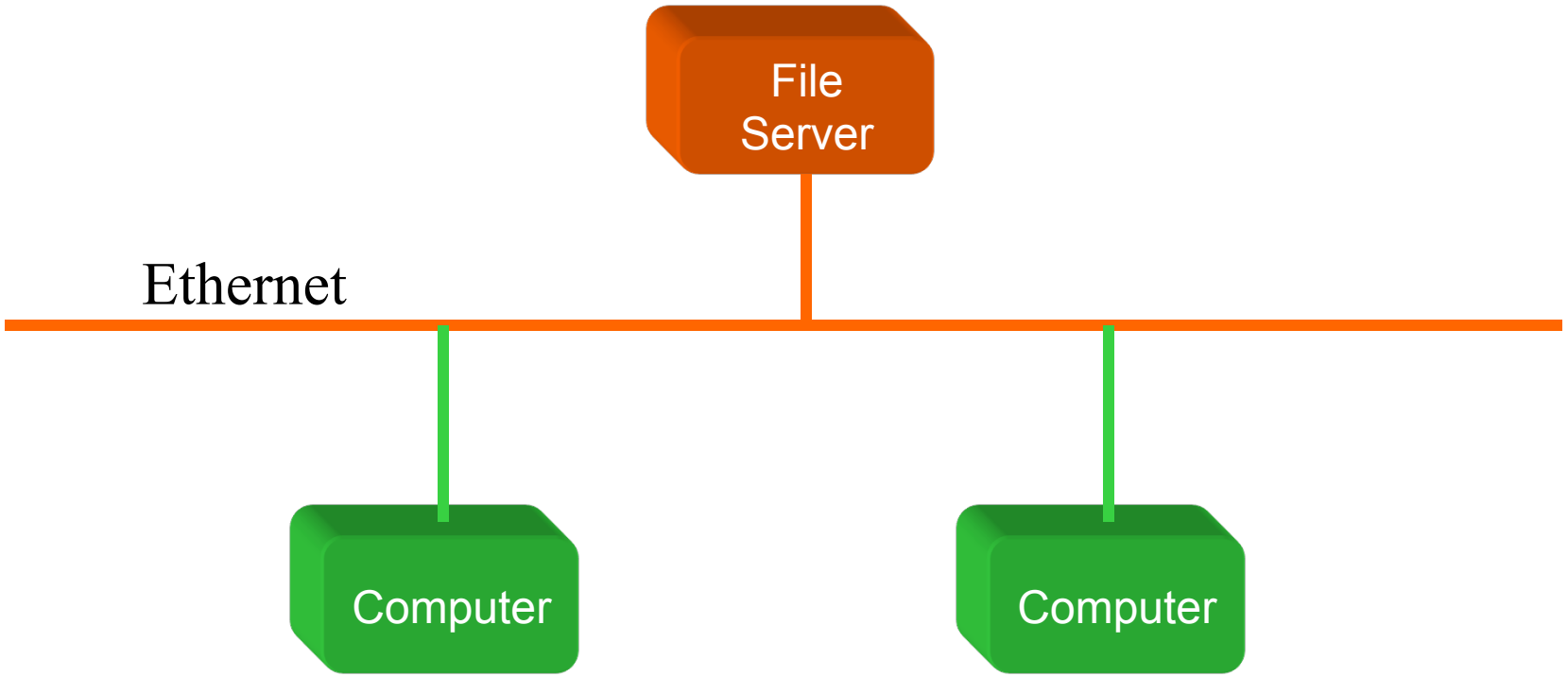


NFS and CIFS



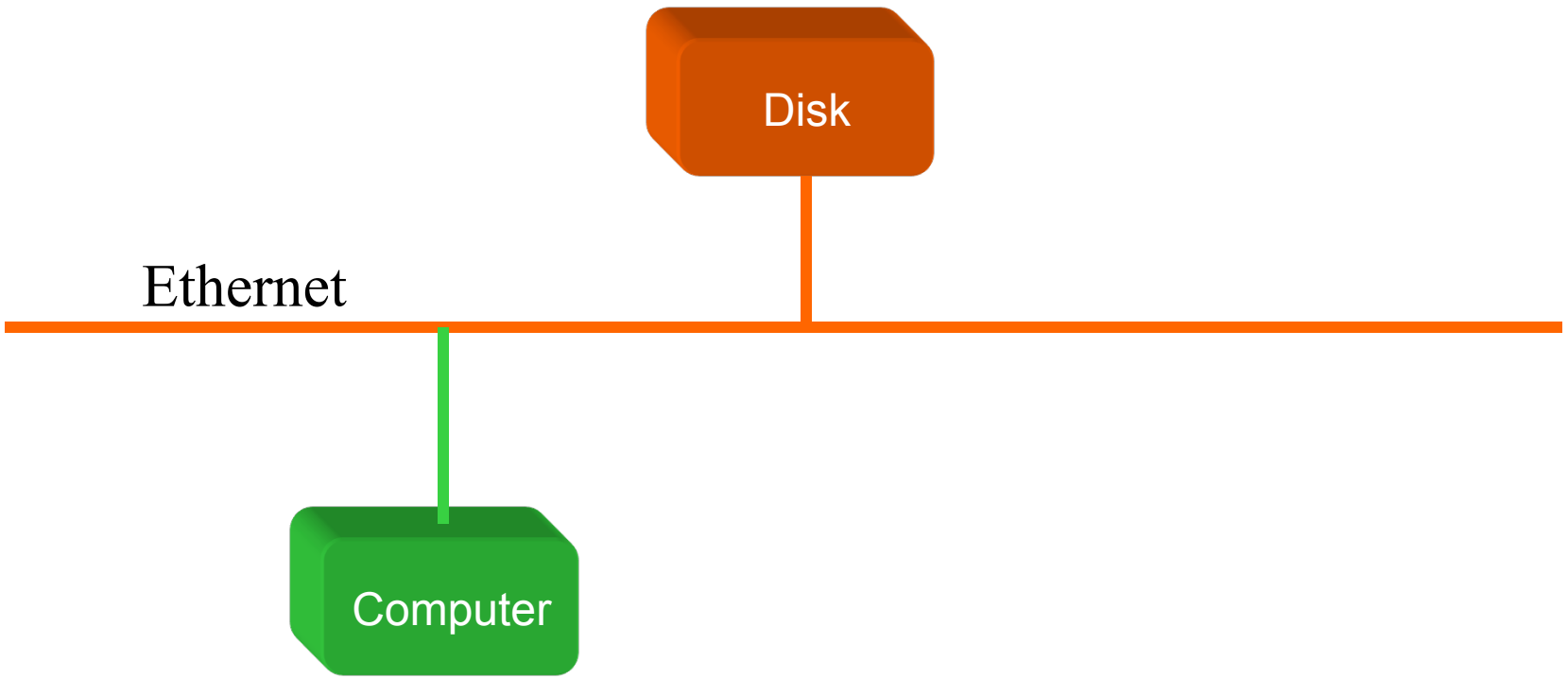
file sharing

Network Attached Storage (NAS)



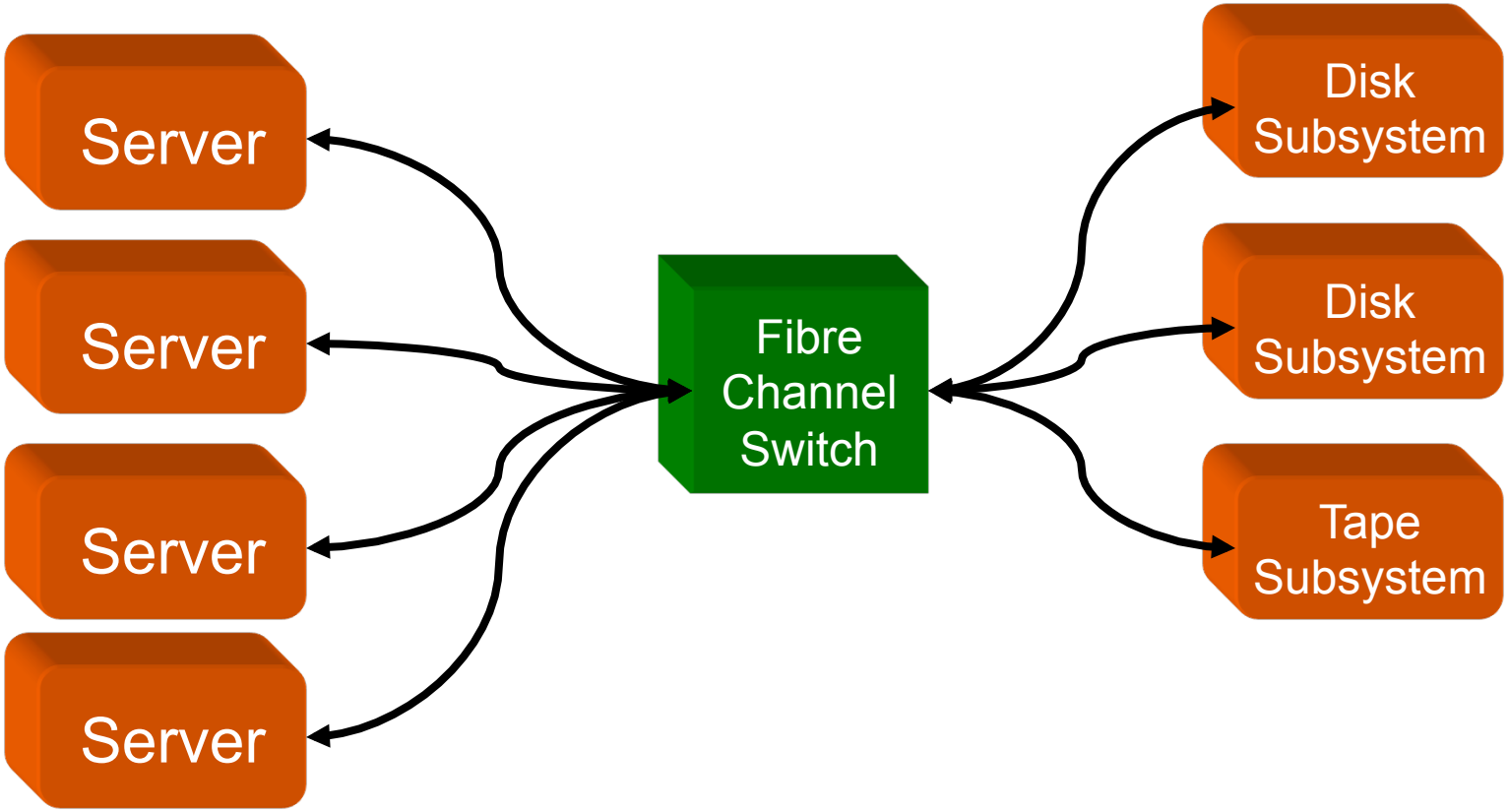
it's just NFS file sharing

iSCSI

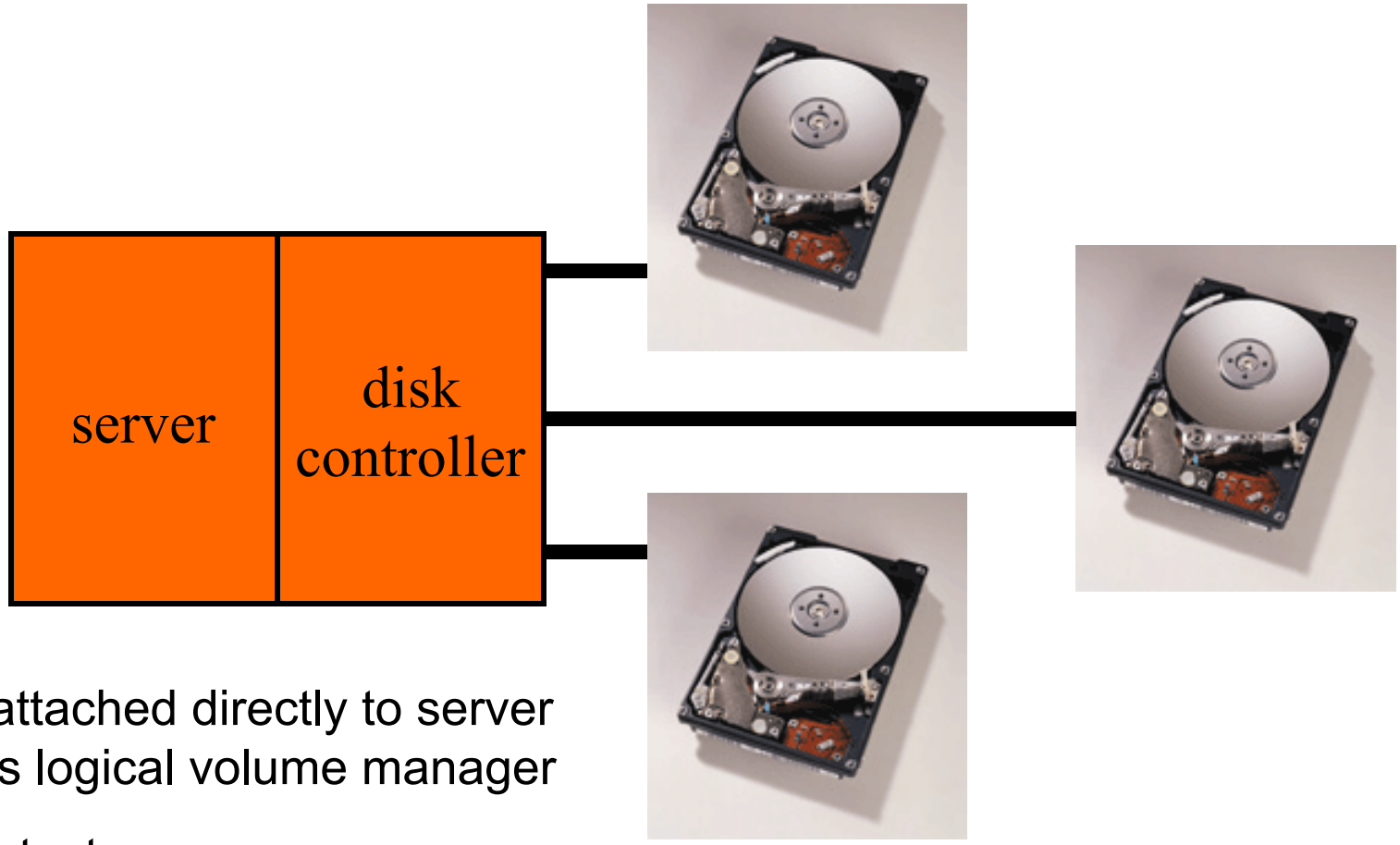


disk block i/o over ethernet

Storage-Area Networks (SAN)



JBOD - Just A Bunch Of Disks



Disks attached directly to server
Use o/s logical volume manager

The fastest

device configuration: decisions

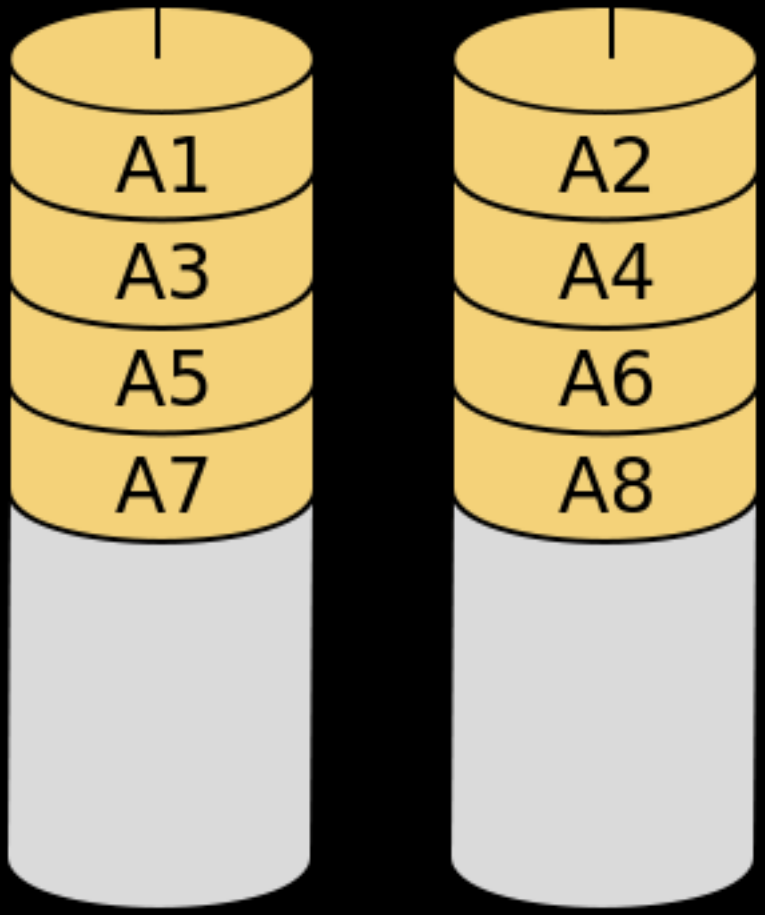
RAID

RAID
Redundant Arrays of Independent Disks

raid diagrams are from wikipedia

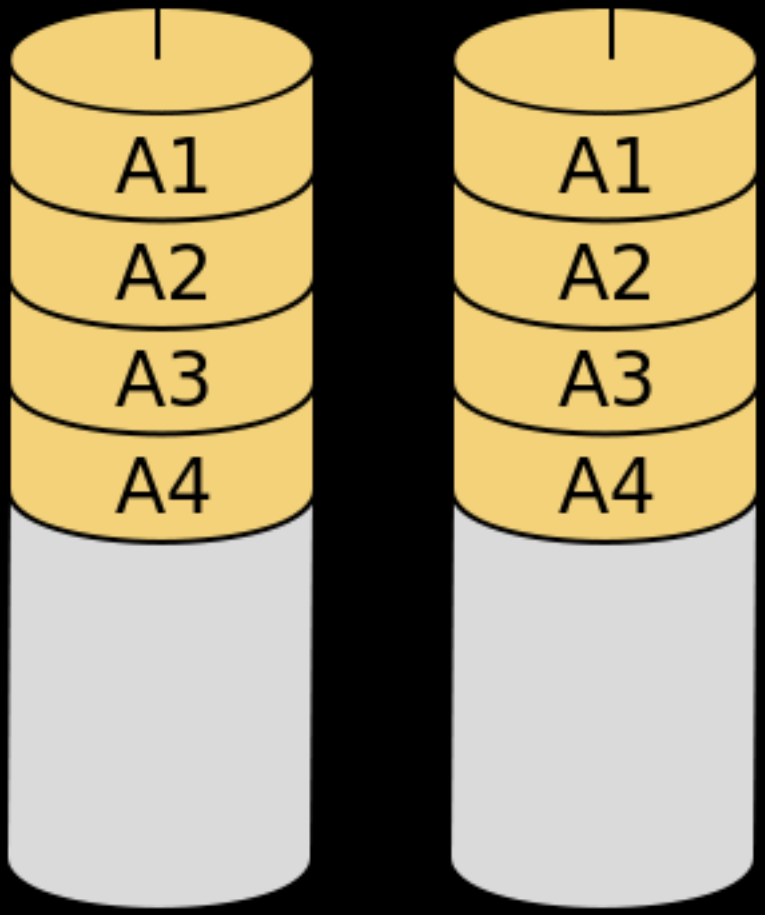
RAID 0: block striping

performance but NO reliability



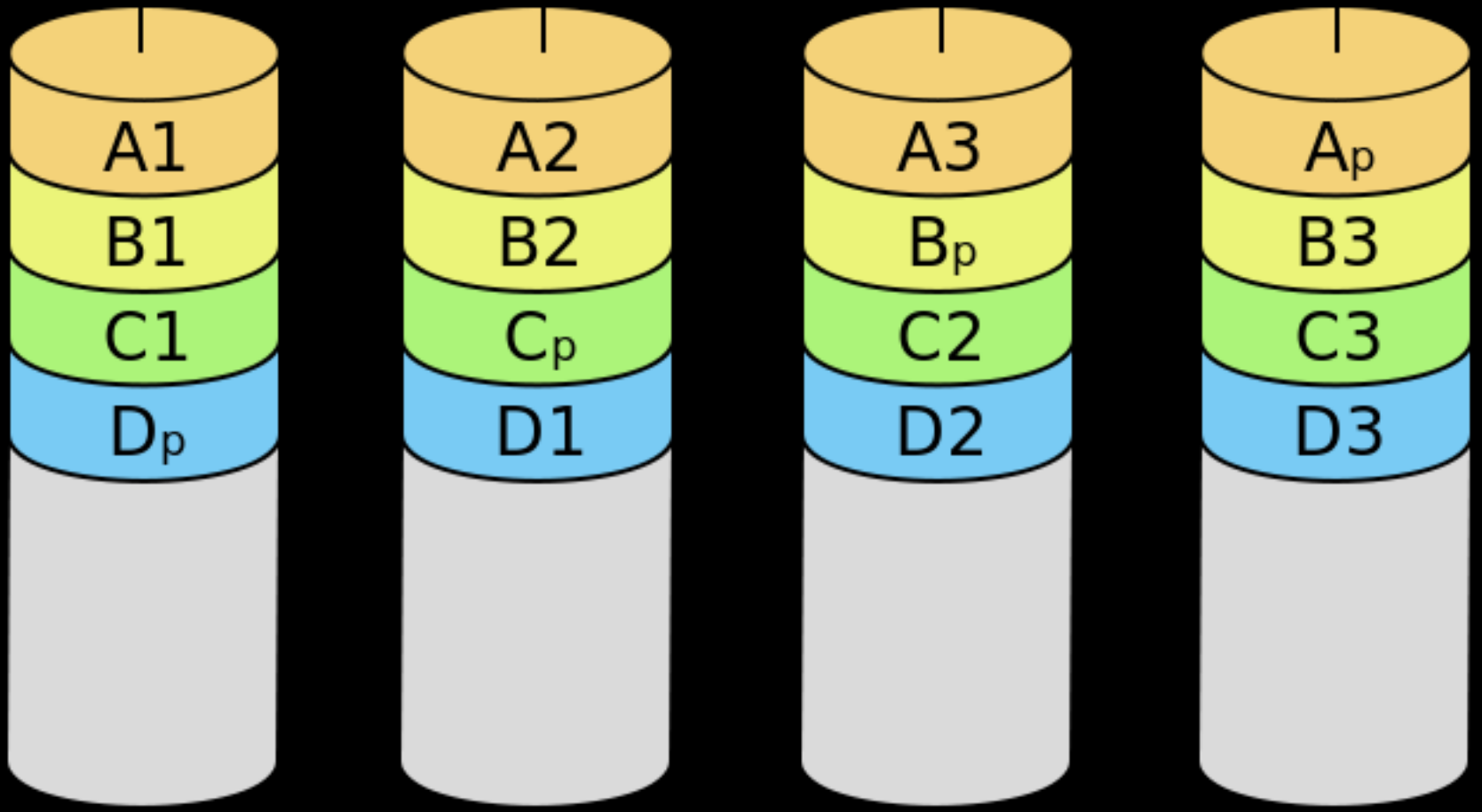
RAID 1: disk mirroring

reliability – two copies



RAID 5: block striping with parity

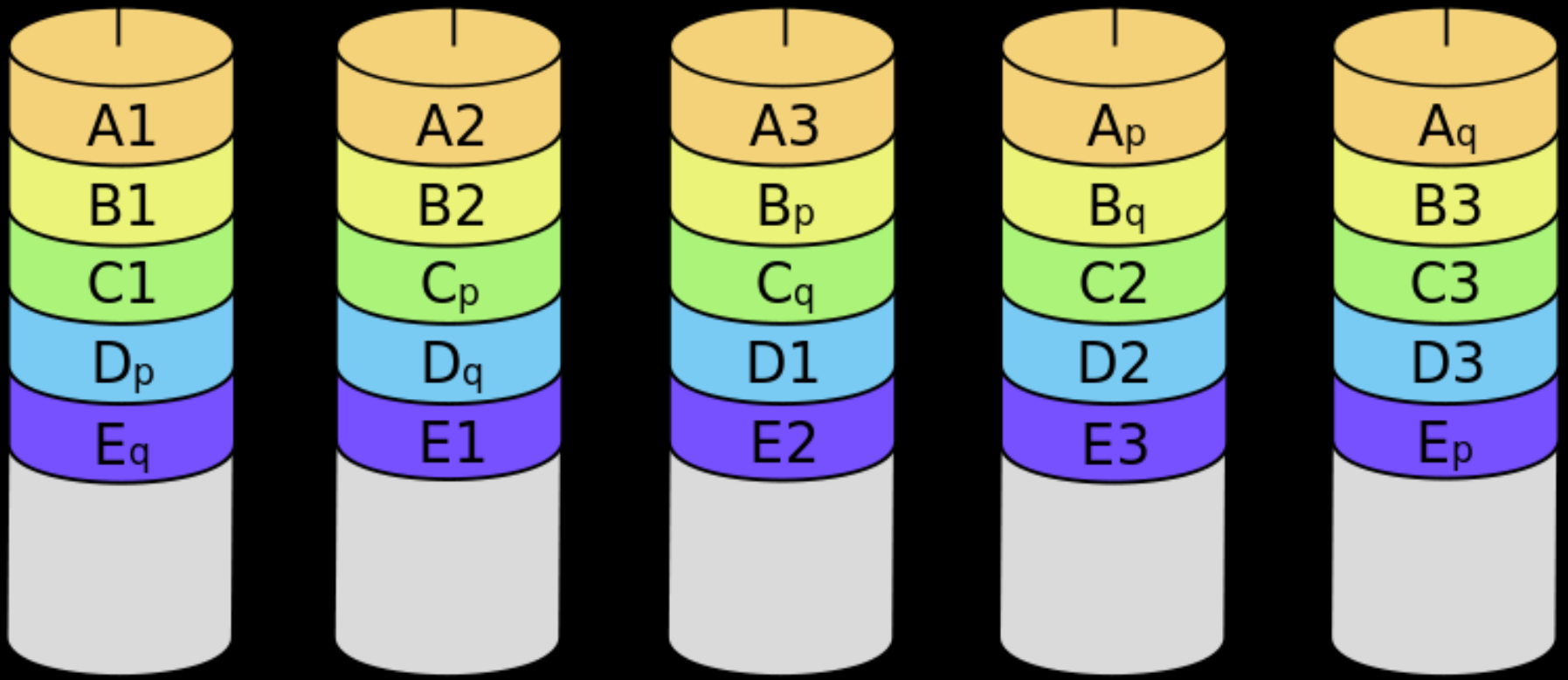
reliability and bad performance



all writes must update 2 drives

RAID 6: block striping with two parity disks

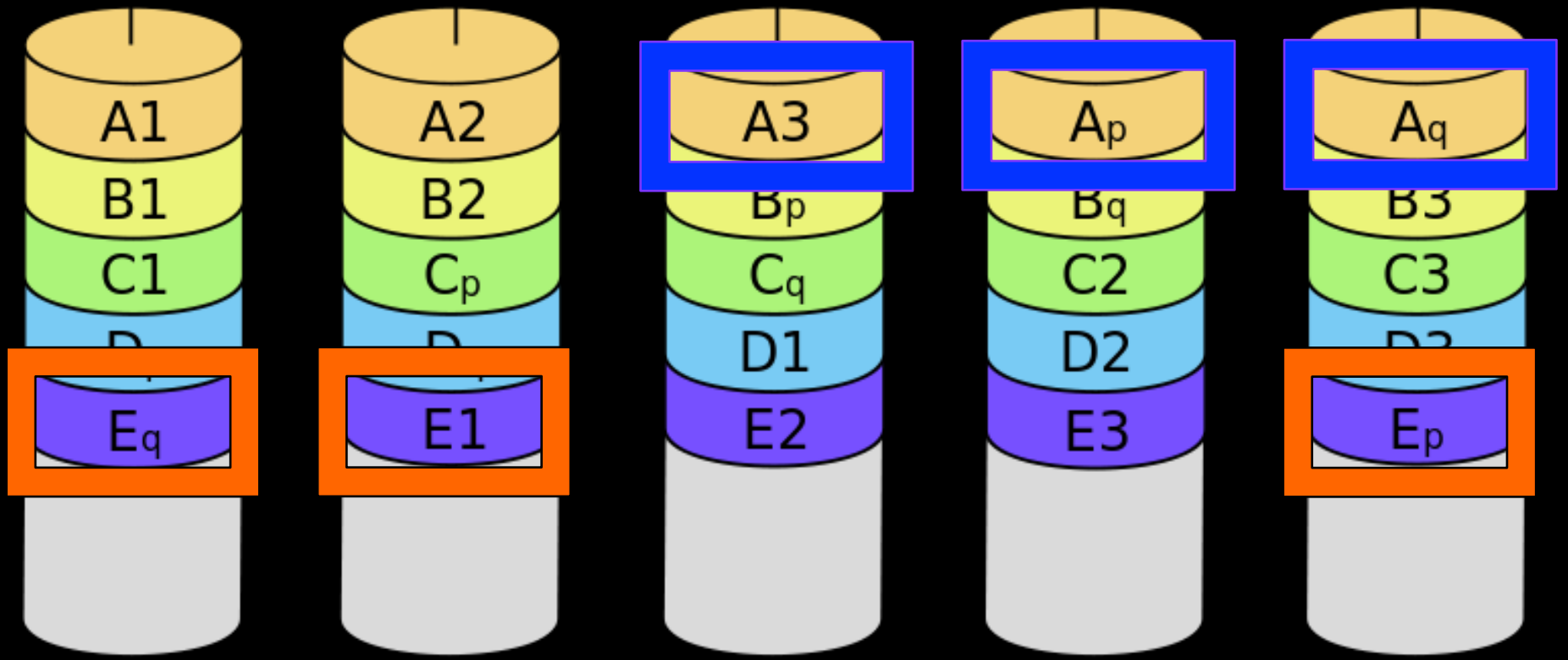
reliability and worse performance



all writes must update 3 drives

RAID 6: block striping with two parity disks

reliability and worse performance



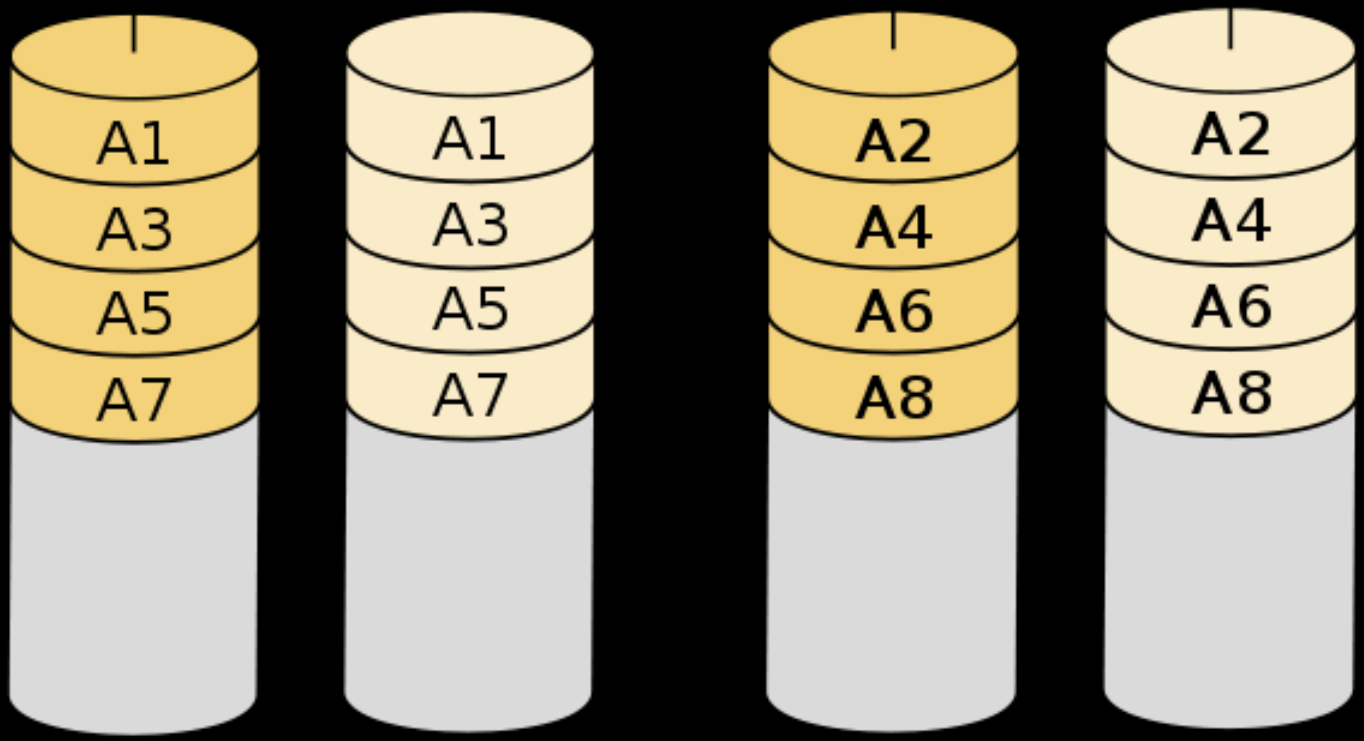
all writes must update 3 drives

write to block E1 has to wait for write to block A3

RAID 10: disk mirroring and block striping

reliability – two copies

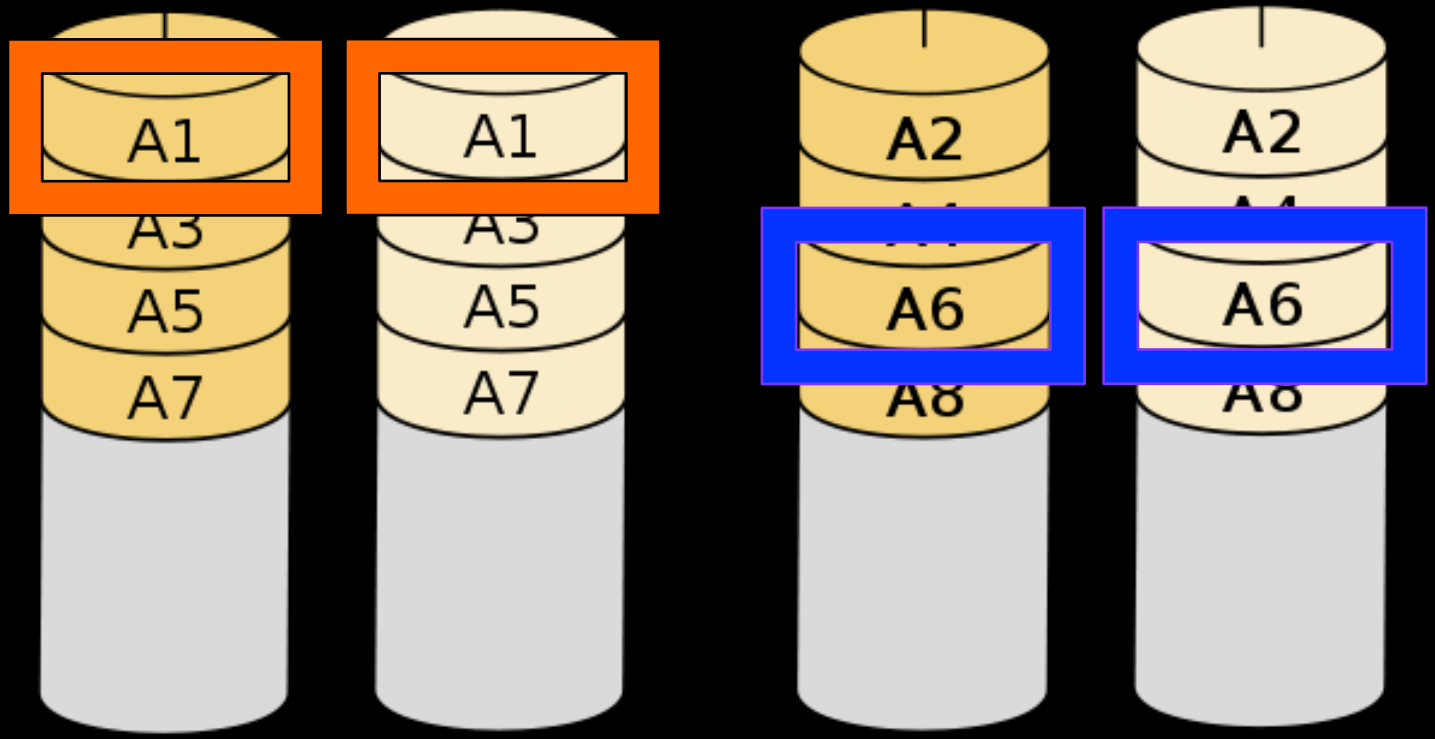
performance – data spread over multiple drives



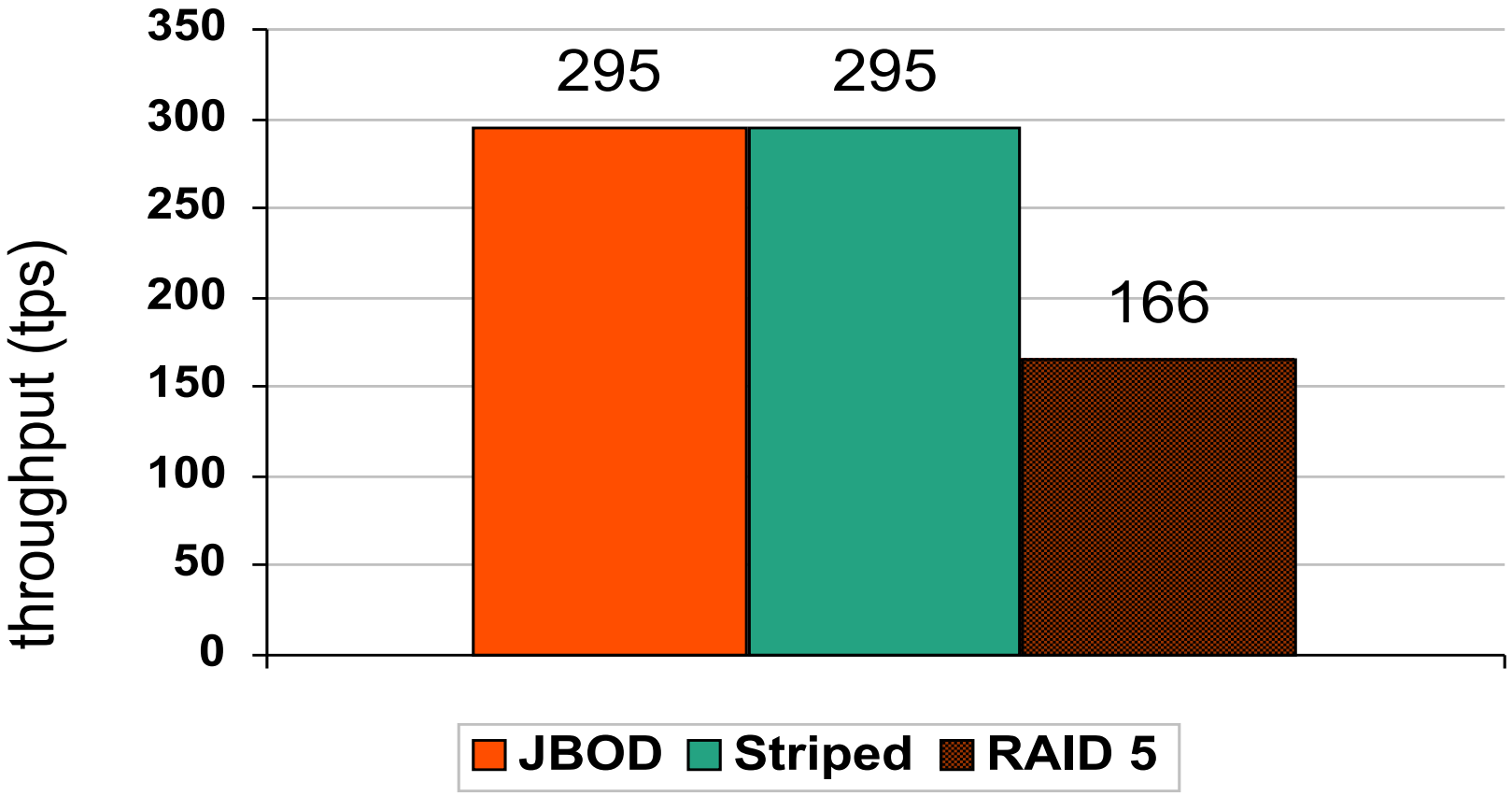
RAID 10: disk mirroring and block striping

reliability – two copies

performance – data spread over multiple drives



Measured
RAID 5 Performance



RAID choices

Type	Description	Use ?
RAID 0	Block striping (no redundancy at all)	Bad
RAID 1	Mirroring	OK
RAID 10	Block striping + mirroring	Excellent
RAID 2	Bit level striping, dedicated parity	Bad
RAID 3	Byte level striping, dedicated parity	Bad
RAID 4	Block striping, dedicated parity	Bad
RAID 5	Block striping with striped parity	Poor
RAID 6	Block striping with dual striped parity	Poor
RAID 60, 6+, DP, etc.	Marketing	Poor

RAID choices – only 1 good one

Type	Description	Use ?
~~RAID 0~~	~~Block striping (no redundancy at all)~~	~~Bad~~
~~RAID 1~~	~~Mirroring~~	~~OK~~
RAID 10	Block striping + mirroring	Excellent
~~RAID 2~~	~~Bit level striping, dedicated parity~~	~~Bad~~
~~RAID 3~~	~~Byte level striping, dedicated parity~~	~~Bad~~
~~RAID 4~~	~~Block striping, dedicated parity~~	~~Bad~~
~~RAID 5~~	~~Block striping with striped parity~~	~~Poor~~
~~RAID 6~~	~~Block striping with dual striped parity~~	~~Poor~~
~~RAID 60, 6+, DP, etc.~~	~~Marketing~~	~~Poor~~

Advancements in technology can never make a silk purse from the RAID 5 / 6 sow's ear. Vendors can't fool mother nature !!!

Local disks will beat SAN storage

QUIZ !!!!

Where are the disk configuration data of your disk array stored ???

How do you recover them when a failure happens ???

Not just the database and transaction logs . . .

- Current backup
- Earlier backups: daily, weekly, monthly
- Archived ai logs: enough to roll forward from your earliest backup
- Binary dump files: dbanalys data size + 15%
- Index rebuild scratch space
- -T : 4GL runtime temp files
- _sqlsrv2 temp files
- Application generated files

OpenEdge RDBMS configuration: decisions

storage area type

maximum-rows-per-block

create limit

toss limit

-recspacesearchdepth

bi file

ai files

Data area types compared

Attribute	Type I Data Area	Type II Data Area
Allocation unit	1 block	1 cluster
Allocation structures	1 per area	1 per object
Allocation concurrency	Low	High
Mixed row data blocks	Yes	No
Mixed Index data blocks	No	No
Fast table/index delete	No	Yes
Table scan	No	Yes
Table-level create limit	No	Yes
Table-level toss limit	No	Yes
Encrypt by table	No	Yes
Multitenant tables	No	Yes
Table partitioning	No	Yes
Maintenance performance	Slow	Fast
Should you use ?	NO	YES !

Tom Bascom says:

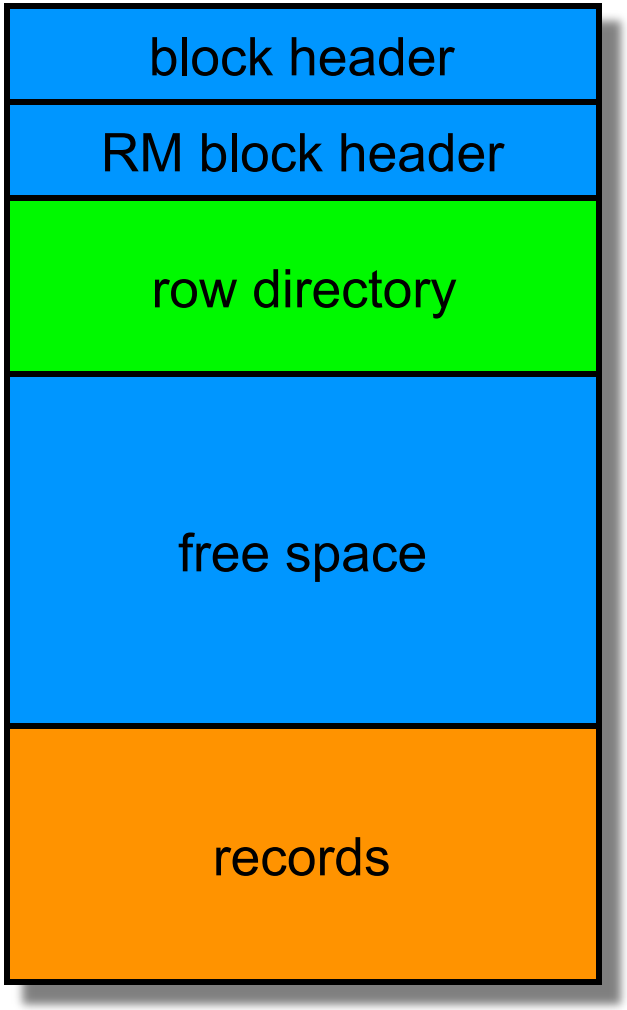
Set Rows Per Block Optimally

	BlkSz	RPB	Blocks	Disk (KB)	Waste/ Blk	%Used	Actual RPB	IO/1,000 Recs
	1	4	3,015	3,015	124	86%	3	333
	4	4	2,500	10,000	2,965	23%	4	250
	4	8	1,250	5,000	2,075	46%	8	125
Original	**4**	**16**	**627**	**2,508**	**295**	**92%**	**16**	**62**
	4	32	596	2,384	112	97%	17	59
	8	4	2,500	20,000	7,060	11%	4	250
Oops!	8	16	625	**5,000**	4,383	44.76	16	62
	8	32	313	2,504	806	90%	32	31
Suggested	**8**	**64**	**286**	**2,288**	**114**	**98%**	**35**	**29**
	8	128	285	2,280	109	98%	35	29

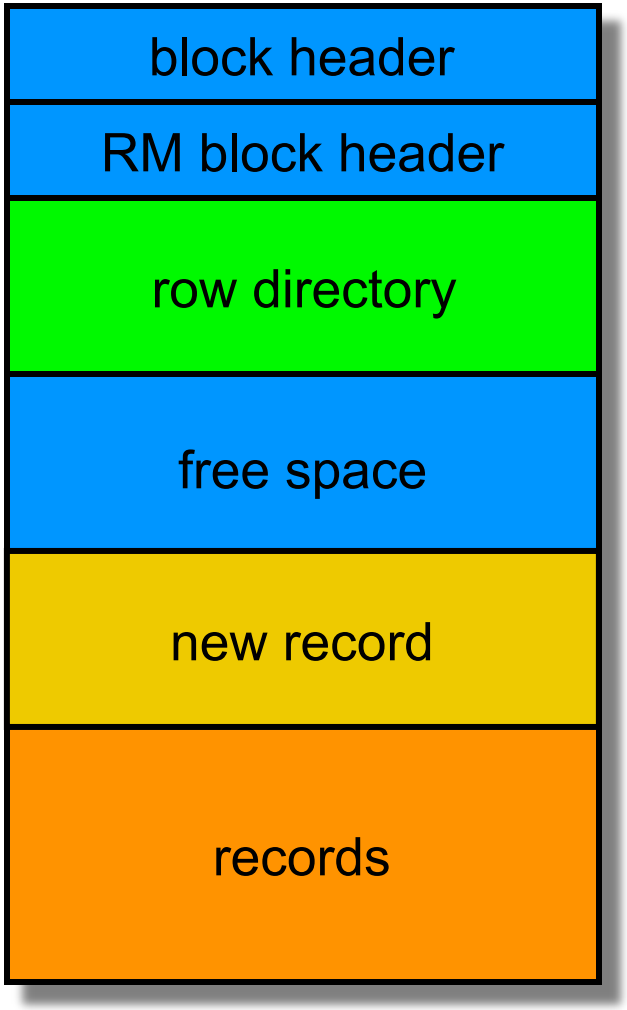
maximum rows per block

- use 64 or 128

Record (RM) data blocks



Record (RM) data blocks

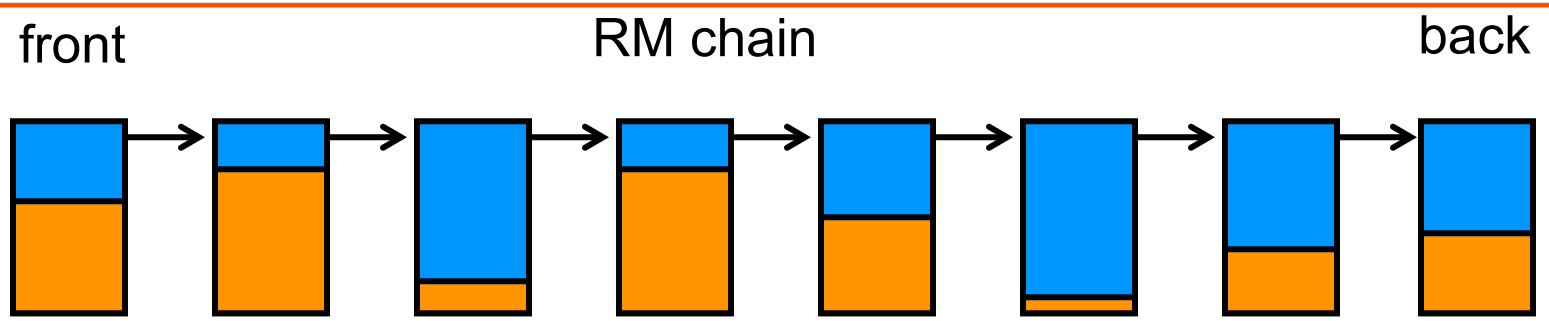


at least create limit bytes

create limit

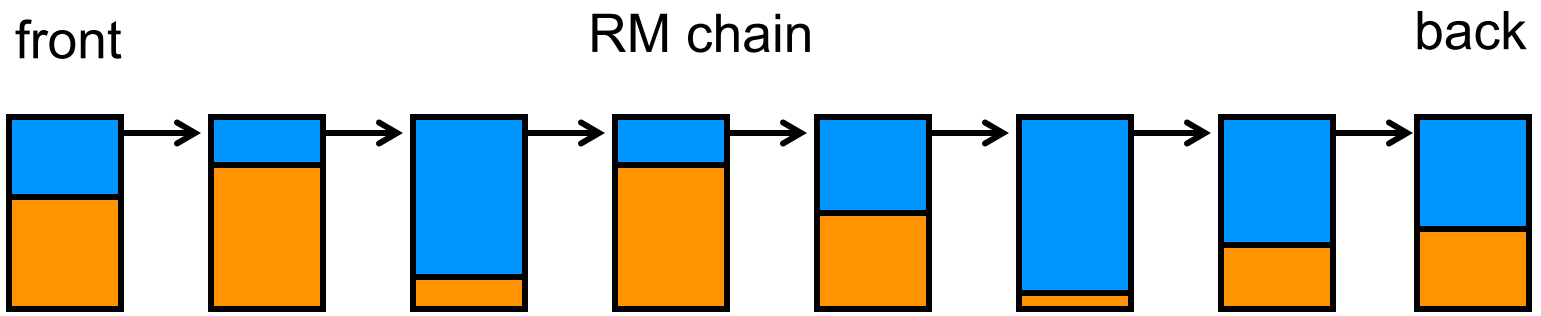
- Change only if sure – you usually don't save much space by optimizing
- Reduce create limit for tables (or LOB columns) whose rows never expand after creation (less wasted space)
- Increase create limit for tables (or LOB columns) that expand after creation
- Must be less than toss limit
- Default value is
  - 75 for 1 and 2 k blocks
  - 150 for 4 and 8 k blocks

toss limit



- Toss limit must be set
  - higher than create limit
  - smaller than average row size
- Default is
  - 150 for 1 and 2 k blocks
  - 300 for 4 and 8k blocks
- Increase when avg row size > toss limit
- Decrease when avg row size < toss limit
- Think: are you wasting your time?

-recspacesearchdepth



search for space begins at first block of RM chain

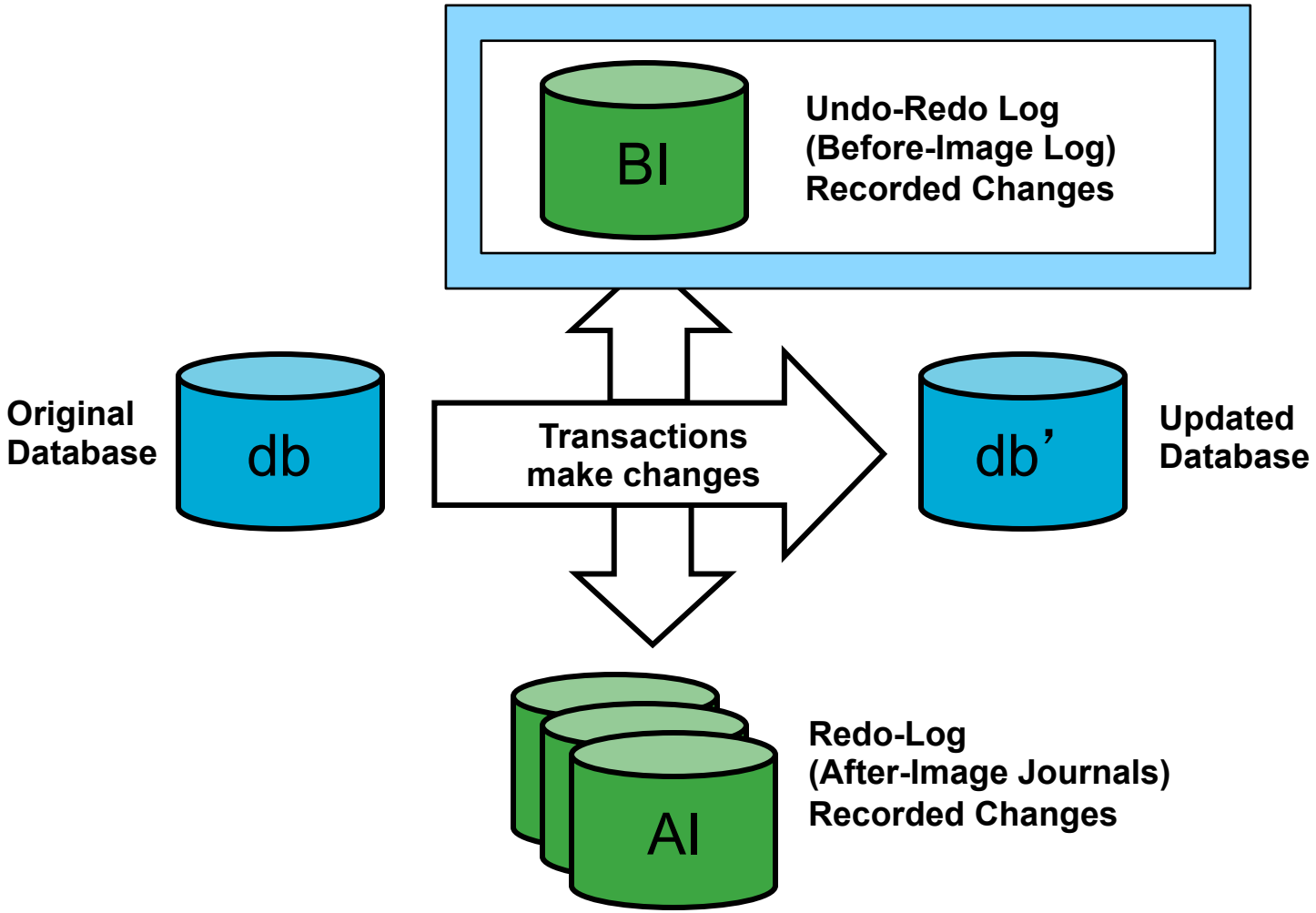
What: -respacesearchdepth parameter determines how many blocks to move from front to back of chain

UNITS: tenths of cluster size, e.g. 9/10

Default setting in 10.2B is 5 (5/10 or 50%)

We believe you should NOT have to adjust this

before-image log



The so-called "Before-Image" file is a lie.

- Does not contain "before images"
- It has a record of all *recent* database changes
- The data are whatever is needed to:
  - Undo or roll back transactions
  - Perform crash recovery
- What is needed depends on the specific operation
  - Row create has "after" row contents – "before" was nothing
  - Row delete has *current* row contents

BI log settings

Parameter	Description	Setting
-bi nnnn	BI cluster size	Enterprise DB: Checkpoints are 2 min long or more (4 mb or higher may be good). Must use APW's. Workgroup DB: Do not increase, smaller is better – maybe 128k
-biblocksize nnn	BI block size	8192 or 16384
-bibufs nn	Number of bi buffers	25 'ish. Smaller than cluster size. Never higher than 100.
BIW	writes filled bi buffers to disk	Always use
-Mf	delayed bi write	Do NOT use

How many BI clusters exist?

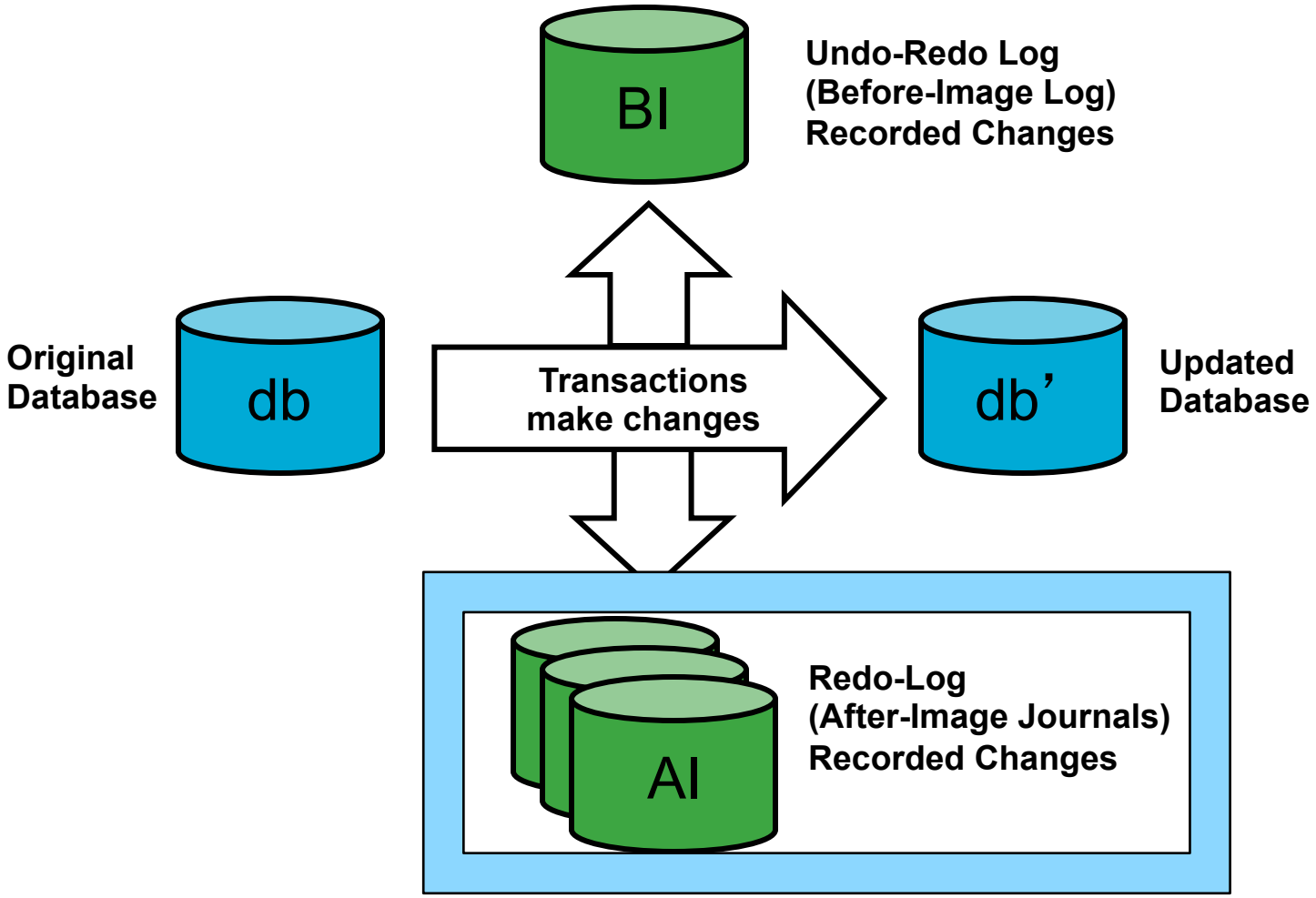
```
find _AreaStatus where
     _AreaStatus-Areanum = 3.

find _dbStatus

display _AreaStatus-Hiwater *
        _dbStatus._DbStatus-BiBlkSize /
        _dbStatus-BiClSize /
        1024
        .
```

Can't tell how many are active though

after-image log



The so-called "After-Image" file is a lie.

- Does not contain "after images"
- It has a record of *all* database changes after a full backup
  - call this point "time 0"
- The data are sufficient to:
  - Recover or recreate everything that happened since time 0
  - Recover or recreate the before-image log
  - Undo or roll back transactions
  - Perform crash recovery
  - Replicate the database in real-time (or later)

ai extents

- Compute amount of ai data by
  - BI cluster size * number of checkpoints per day, week, month
- minimum 4 extents
  - current - new notes go here
  - next 1 - used when current fills
  - next 2 - in case next 1 fills
  - full - used to be current, needs archiving

AI Log settings

Parameter	Description	Setting
number of ai extents		minimum 4, better is 9 time to fill all ai extents determines number and size
-aiblocksize nnn	AI block size	8192 or 16384, same as BI block size
-aibufs nn	Number of bi buffers	25 'ish. Smaller than cluster size. Never higher than 100.
AIW	writes ai buffers to disk	Always use
ai daemon	manages and archives filled extents	Always use

What should you do?

To do list

- always use type ii data areas
- configured properly
- use RAID 10 with direct-attached disks
- use RAID 10 with SAN if direct-attached not possible
- use SSD when possible
- have spares for SSD devices
- have spares for spinning rust devices
- use apw, biw, aiw auxiliary processes
- use after-image journalling
- use ai archive daemon
- remember all the other stuff you need space for

* YMMV, **mistakes**, transportation, meals, and accomodations not included

That’s all we have time for today, except

Answers

Email:

gus@bravepoint.com