RAID - On a Budget

MUUG – January 11, 2004
John Schulz

Overview

- Background
- Disk Technologies
- RAID Terminology
- RAID Technologies
- Examples
- Some Numbers

Background

- Based on my personal experience and mainly "non-scientific" findings
- Talk will focus on entry level options
- Two main goals:
  - Lowest cost per MB possible with redundancy we can live with
  - A bit better performance but still remembering cost objective

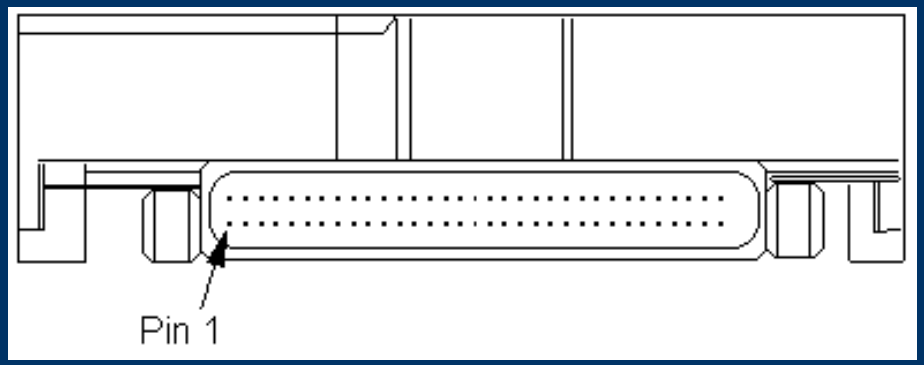
Disk Technologies

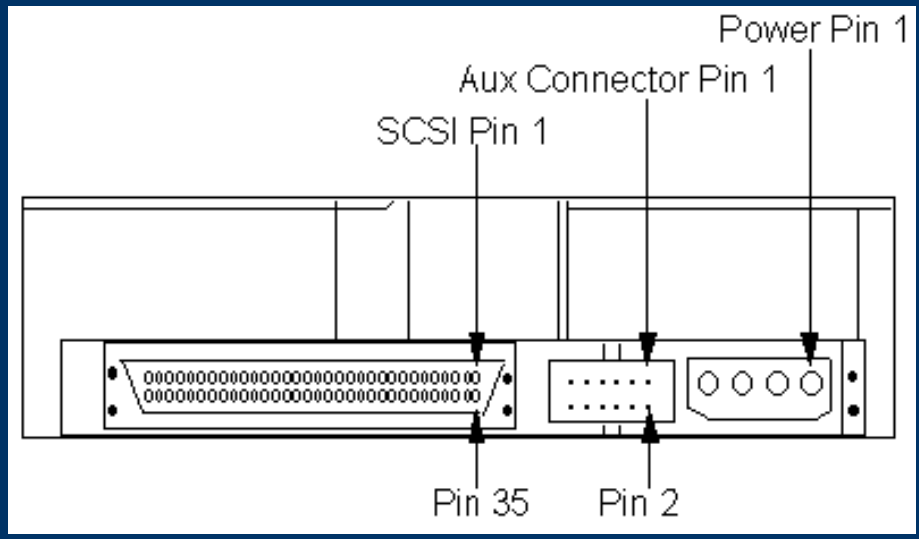
Disk Technologies - SCSI

- SCSI - Small Computer System Interface
- Up to 320MB/sec, 10K, 15K RPM drives
- Up to 300GB drives available
- Low CPU load
- Optimized for multitasking applications
- Up to 16 devices on single wide (16 bit) bus
- Highest cost per MB

Disk Technologies - SCSI

- Connectors:
  - 68 pin with separate power, set SCSI ID on drive
  - 80 pin (SCA) connector with integrated power, set SCSI ID external to drive





Disk Technologies - SCSI

- Pricing:
  - 36GB 10K @$220 ($6.10/GB)
  - 73GB 10K @$350 ($4.80/GB)
  - 73GB 15k @ $638 ($8.70/GB)
  - 146GB 10K @ $714 ($4.90/GB)
  - 300GB 10K @ $1,800 ($6.00/GB)

Disk Technologies - PATA

- PATA - Parallel Advanced Technology Attachment
  - What we all know as IDE
- Up to 133MB/sec
- Up to 400GB drives
- 2 drives per channel (Master/Slave)
- Issues with older BIOS support for drives over 128GB

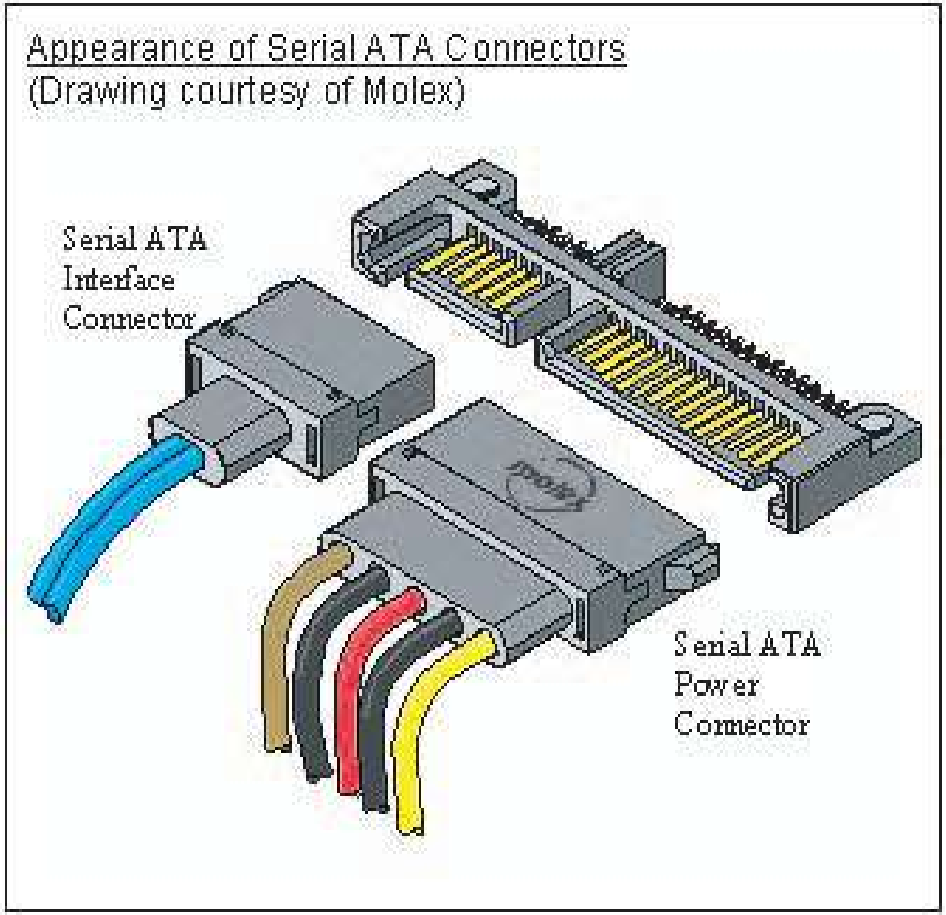
Disk Technologies - SATA

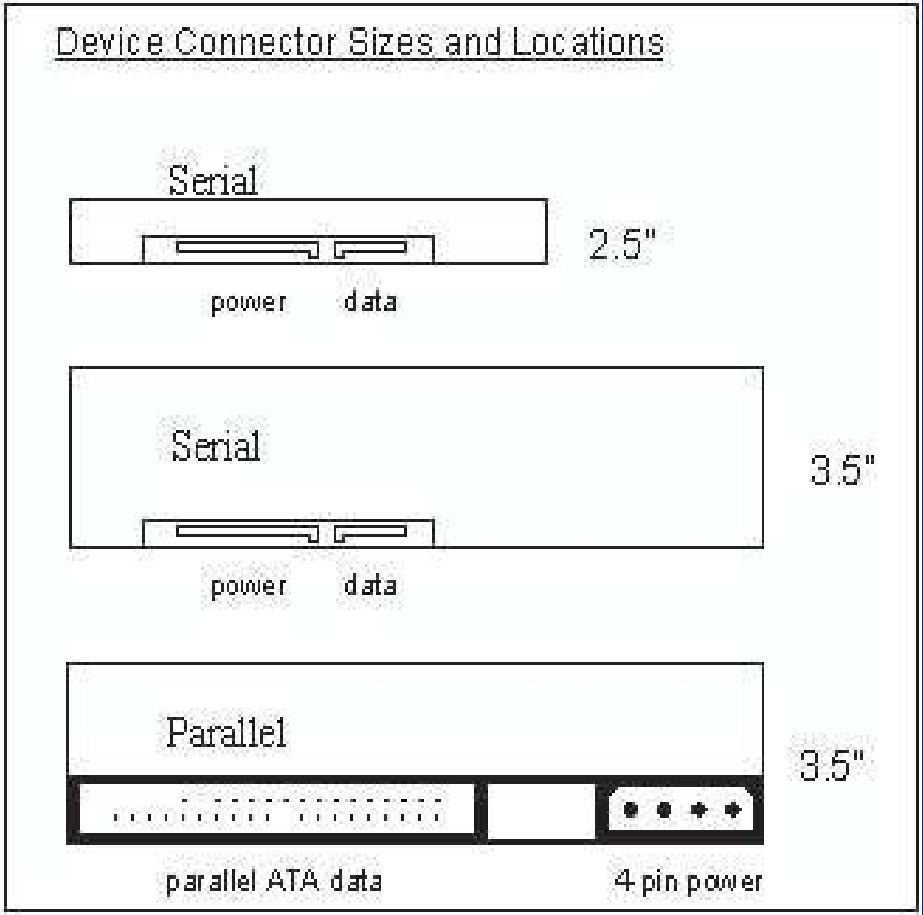
- SATA - Serial Advanced Technology Attachment
- Starting at 150 MB/sec
- 1 drive per channel
- Better performance than PATA
- Different Power connector on newer drives
- Specification supports hot plugging

Disk Technologies - SATA

- Replacing PATA
- Was more expensive than PATA but now about the same cost
- Pricing (7200 RPM):
  - 80GB @ $85 ($1.10/GB)
  - 250GB @ $200 ($0.80/GB)
  - 300GB @ $262 ($0.87/GB)
  - 400GB @ $600 ($1.50/GB)

Disk Technologies - SATA





RAID Terminology

RAID Terminology

- Redundant Array of Inexpensive disks
  or
- Redundant Array of Independent disks

- A way of storing the same data in different places (thus, redundantly) on multiple hard disks.
- Goals:
  - Fault tolerance
    - Since multiple disks increases the mean time between failure (MTBF), storing data redundantly also increases fault-tolerance.
  - Performance
    - By placing data on multiple disks, I/O operations can overlap in a balanced way, improving performance.

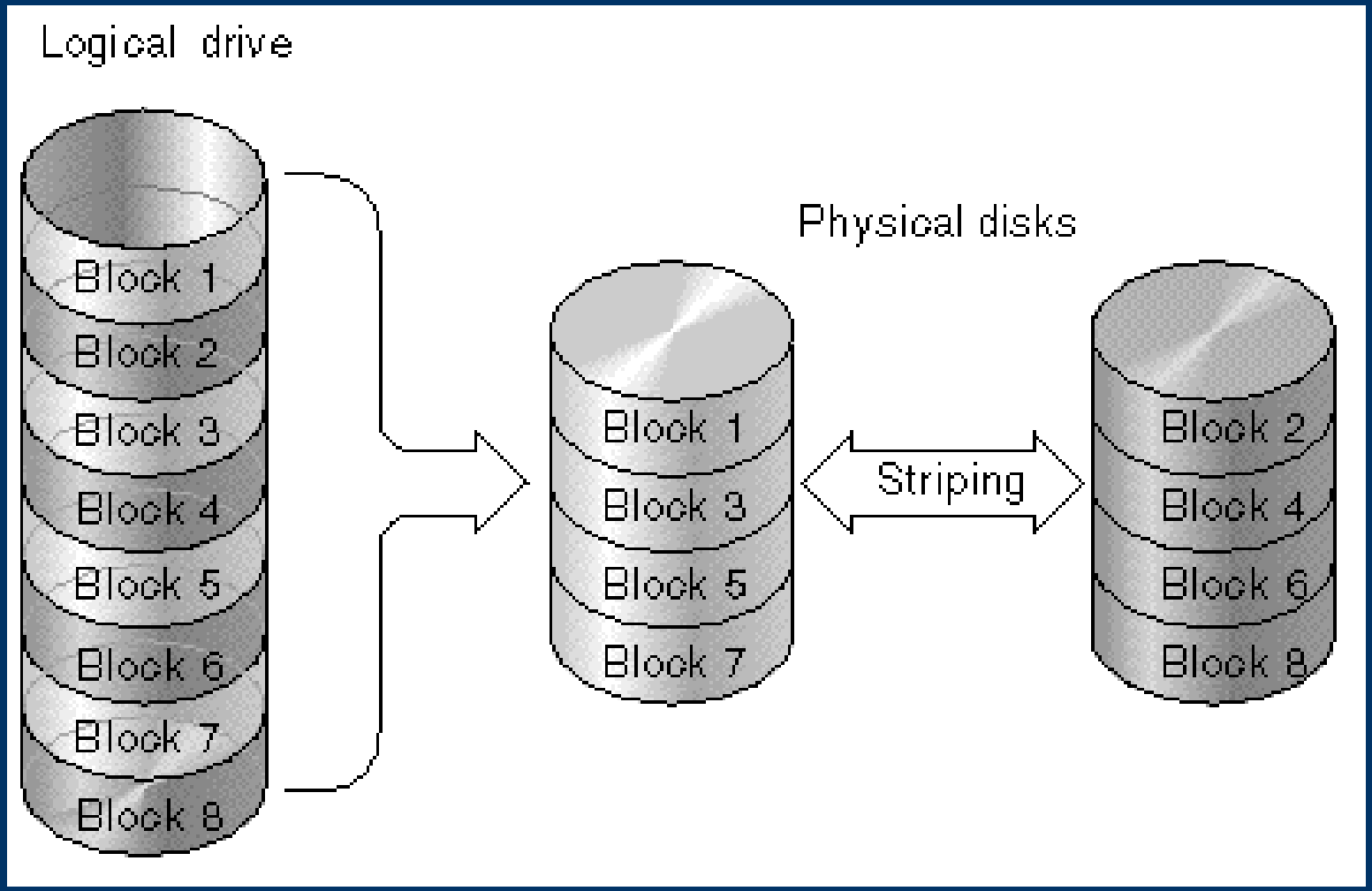
RAID Terminology – JBOD

- Data spanning over a number of disks to create a large volume
- A way to get around Partition size limits on W2K
  - Create a number of 2GB partitions
  - Set them to be dynamic
  - Create a Spanning Volume over the dynamic partitions

RAID Terminology - RAID 0

- RAID 0 - Striping of data
- Data is broken into logical blocks and is striped across several drives
- No redundancy – if 1 drive fails your data is gone
- Best Performance
- Not really RAID – there is no redundancy

RAID Terminology - RAID 0

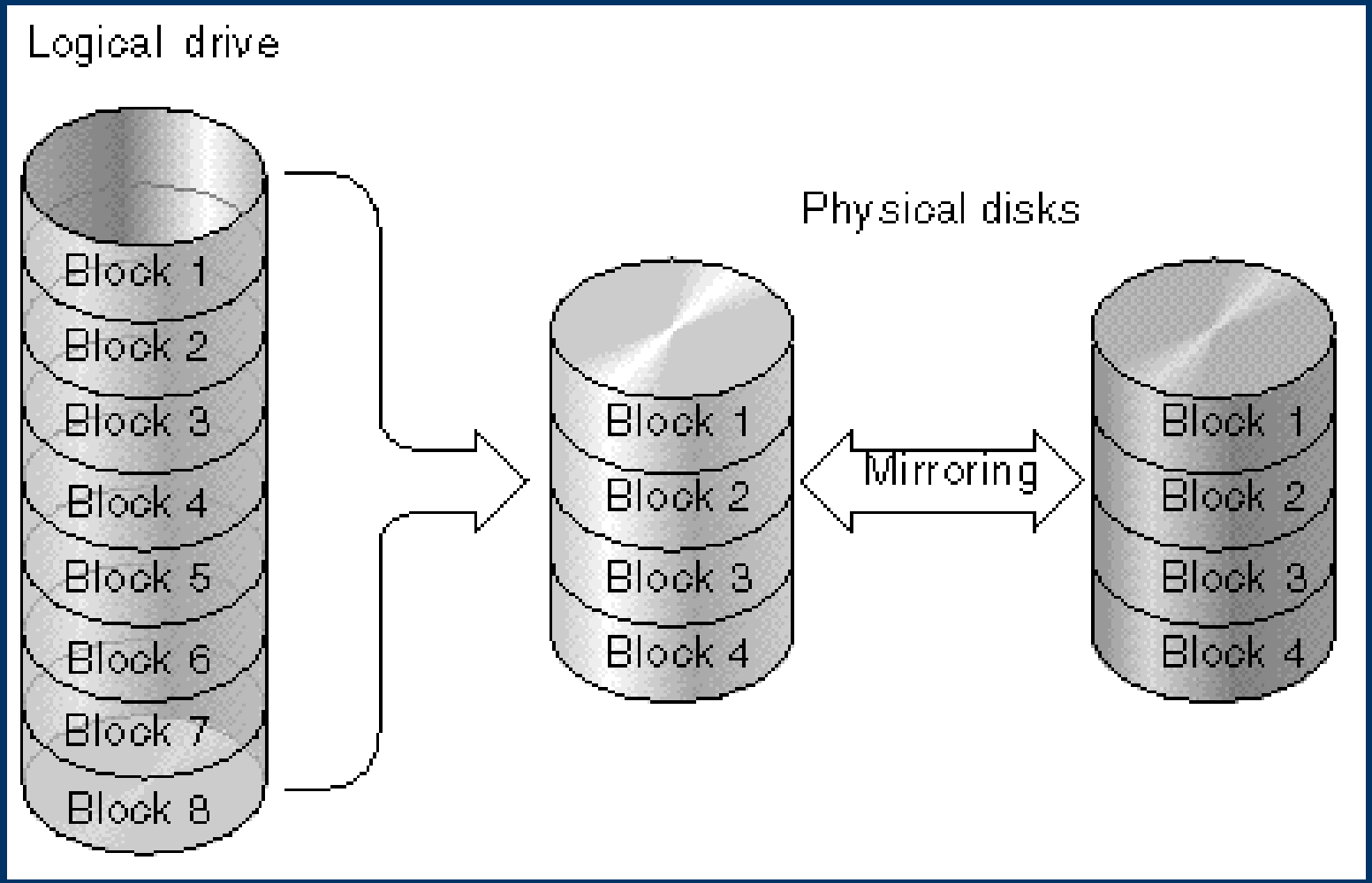


RAID 0 – data stripped over disks

RAID Terminology - RAID 1

- RAID 1 - Mirroring of data
- A copy of the same data is recorded onto two or more drives
- A data read can round robin the drives, improving performance
- Good redundancy – can survive a disk failure
- If drive fails, degraded performance good
- Under Software RAID 1, you can break mirrors to reduce backup downtime windows.

RAID Terminology - RAID 1



RAID 1 – data mirrored over disks

RAID Terminology - RAID 5

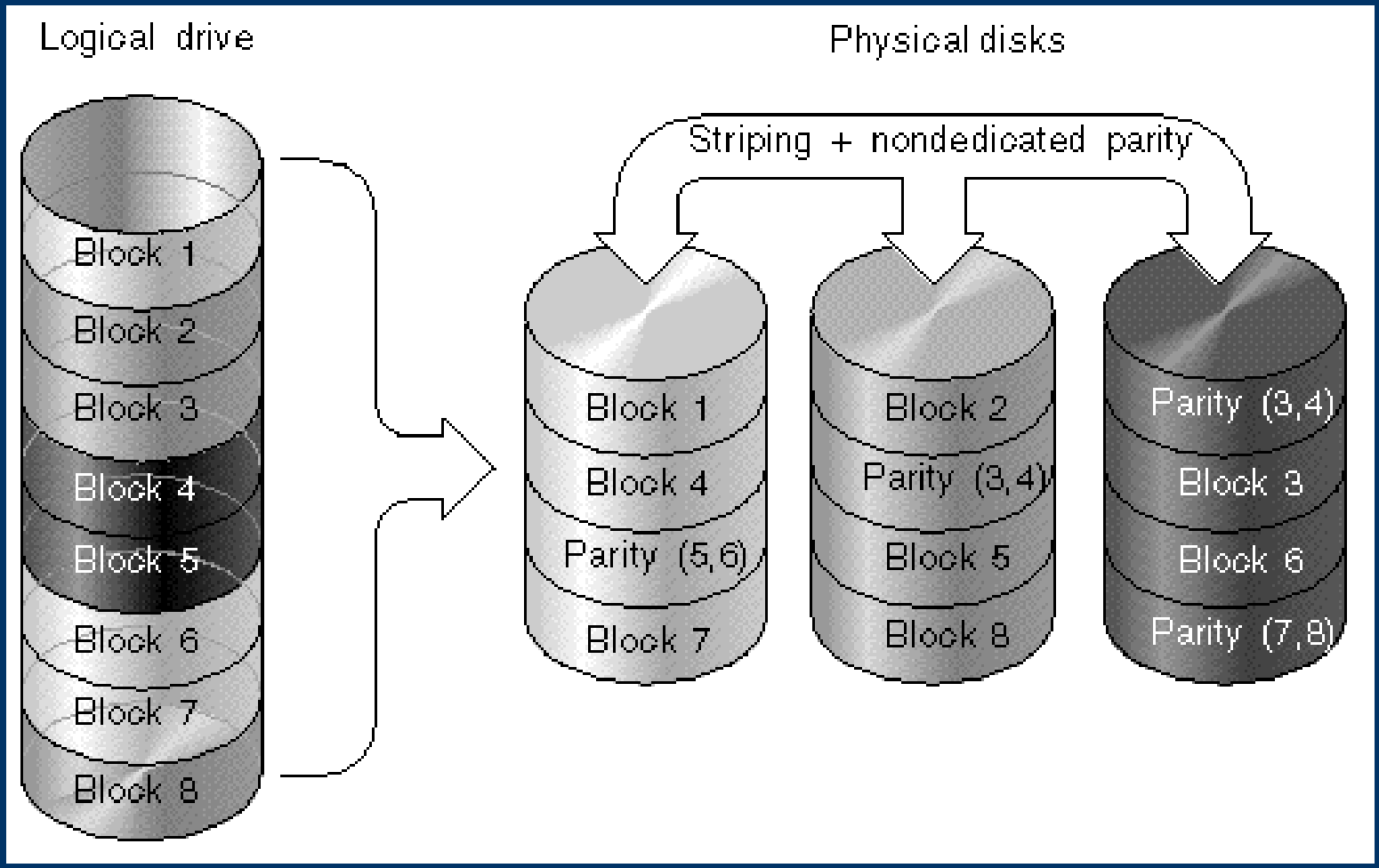
- RAID 5 - Multiple-block striping with distributed parity
- Data and its parity are never stored on the same disk

- In the event that a disk fails, original data can be reconstructed using the parity information and the information on the remaining disks
- Good redundancy – can survive 1 disk failure
- Add a hot spare to minimize degraded performance window

RAID Terminology - RAID 5

- Degraded performance can be bad – time to recalculate missing data
- Degraded performance distinguishes better RAID 5 controllers
  - Good to test degraded performance when testing RAID systems
- RAID 5 Size: (number drives – 2) x drive size
  - 1 hot spare
  - Lose 1 drive of space for parity (actually a bit more)

RAID Terminology - RAID 5



RAID 5 – data and parity spread over disks

RAID Terminology - RAID 10 and 01

- RAID 10: Mirroring and striping
- RAID 01: Striping then mirroring
- Info (using at least 4 drives):
  - Array Capacity: (Size of Smallest Drive) * (Number of Drives) / 2.
  - Storage Efficiency: If all drives are the same size, 50%.
  - Fault Tolerance: Very good for RAID 01; excellent for RAID 10 (you could lose up to 3 drives).
  - Availability: Very good for RAID 01; excellent for RAID 10.
  - Degradation and Rebuilding: Relatively little for RAID 10; can be more substantial for RAID 01.

RAID Technologies

RAID Technologies - Software

- The O/S creates and maintains RAID devices
- Any data copying/calculation performed by the CPU

- Linux: supports RAID 0, 1, 5
  - fd file system and raid utilities
  - Logical volume management
  - Can be configured at install time with Kickstart (Redhat)

- Windows
  - W2K Pro: striping and spanning
  - W2K Server adds RAID 1 and 5

RAID Technologies - Software

- On-board controller:
  - SATA ICH5R chipset
  - Adaptec RAIDHost 39320D-R

- Have BIOS support from the chipset for booting
  - After that all of the RAID functionality is handled via the host OS.

RAID Technologies - Host-based

- Utilizes a Hardware based controller card with embedded RAID functionality
  - O/S works with Hardware card/chipset
  - Hardware card interacts with physical disks
  - Lower end boards have only RAID 0 and RAID 1

- It's all about the drivers
  - O/S must have drivers/software to support the card.
  - Installing on RAID boot partition usually more difficult than creating a data partition

RAID Technologies - Host-based

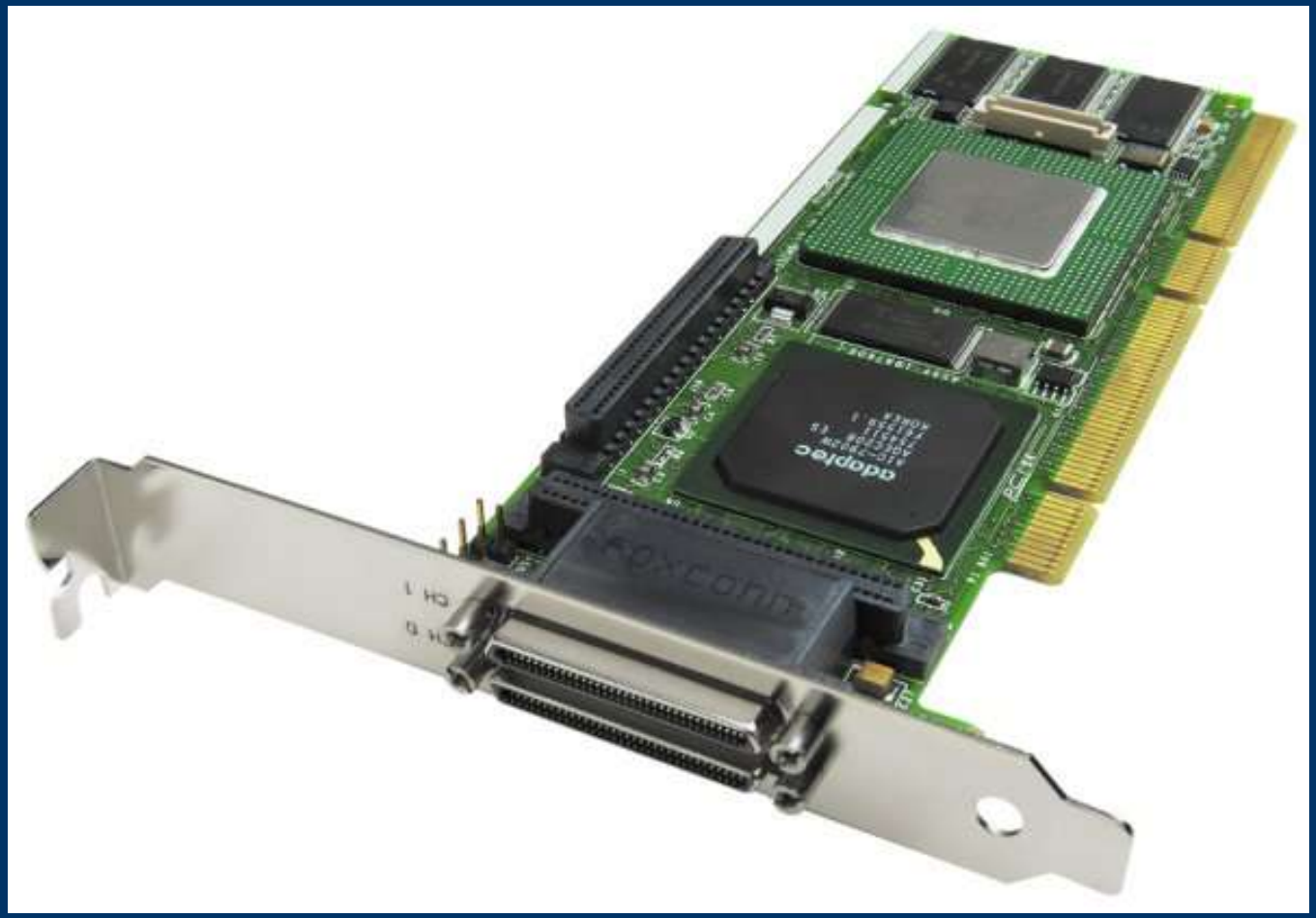
- Card based:
  - PCI based card with drive connectors
  - Drives - can either be in the system or external to the system

- External storage for a Host based controller a step toward fully external RAID.

RAID Technologies - Host-based

- SCSI PCI card:
  - 1 or 2 channel SCSI controller
  - Internal or external SCSI disks
  - Less expensive that fully external system
  - Requires space and power if internal
- SCSI Zero channel card:
  - SCSI controller plugs into motherboard
  - Card adds RAID functionality to existing on-board SCSI controller

RAID Technologies - Host-based



Adaptec 2200 – On-board hardware RAID

RAID Technologies - Host-based

Adaptec 2015S – Zero channel RAID card

RAID Technologies - Host-based

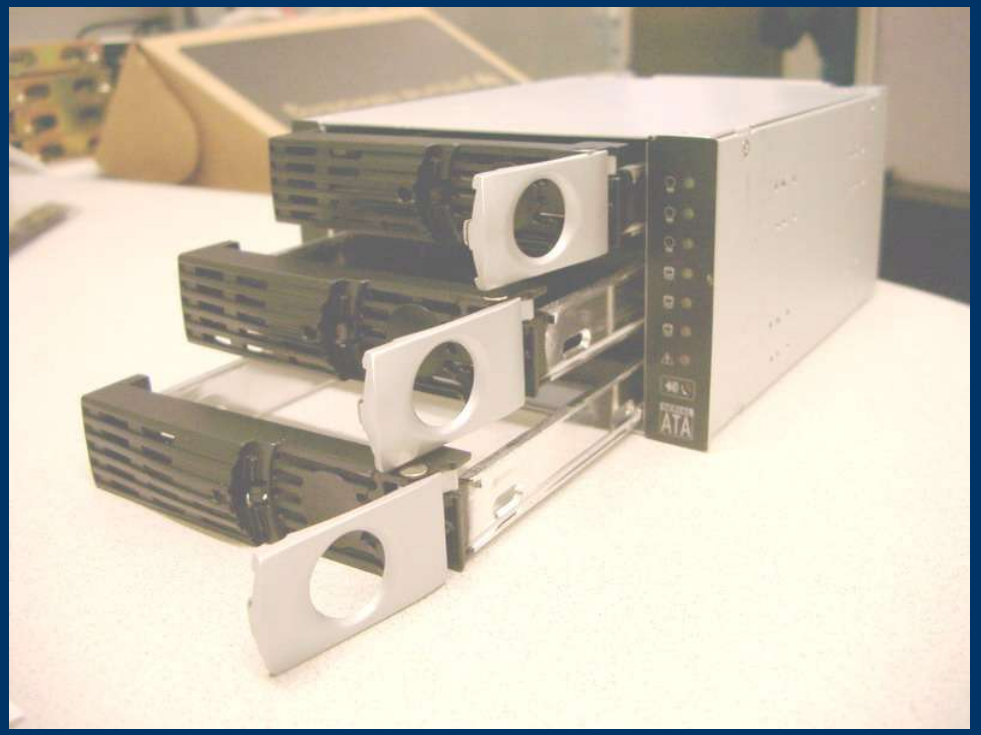
- SATA PCI card:
  - 2, 4, 8, 16 channel cards available
  - Adaptec has a kit with 4 channel controller and 4 drive enclosure

- Various manufactures make SATA drive enclosures
  - Cremax (ICY-Dock)
  - Startech
  - Storcase

RAID Technologies - Host-based

Adaptec 2410 Kit – 4 SATA drives in 3 bays

RAID Technologies - Host-based



Cremax MB235 SPF – 3 drives in 2 bays

RAID Technologies - Host-based

StarTech SATABAY3 - 3 Drive Serial ATA Backplane

RAID Technologies - External Controller

- System and O/S independent
- External interface is SCSI
- Available from many vendors
- General structure:
  - 1 or more external SCSI channels
  - Create a RAID 5 Disk array with Hot Spares
  - Create a number of logical volumes
  - Map each logical volume to a SCSI address using SCSI ID and LUN
- Watch for:
  - For maximum supported file system size
  - Number of internal channels

RAID Technologies - External Controller

- Storcase:
  - SCSI-SCSI, SCSI-FC, and SCSI-SATA
  - Serial User interface for management

  - SCSI:
    - 9 or 14 bay
    - Dual channel SCSI-SCSI enclosure
    - Dual or Single RAID controller
    - Uses 80 Pin drives SCSI drives

  - SATA:
    - 12+ bays
    - Dual channel SCSI-SATA enclosure
    - Dedicated SATA channel per disk

RAID Technologies - External Controller

Storcase dual channel with 9 SCSI drives

RAID Technologies - External Controller

Storcase with 12 SATA drives

RAID Technologies - External Controller

- Nexsan
  - 14 drive bays
  - PATA drives
  - Dedicated PATA channel per disk
  - GUI windows based tools
  - On the net (web based interface)

Nexsan ATAboy2

RAID Technologies - External Controller

Nexsan ATAboy2 - back

Examples

- Basic setup and costs
- Various cost/GB
- Usually trade off between reliability and cost

Example 1: Software RAID

- SuperMicro 5013C-T 1U system @ $1,200
  - 2 x hotswap SATA bays
  - On-board ICH5R
  - 2 x 300GB SATA drives @ $262
    - 10GB for O/S, 290GB for RAID 1 storage
  - $1724/290GB= $5.94/GB

Example 2: SATA Host-based card

- Redhat 7.3 Box with internal SATA
  - 1 x Adaptec 2810SA SATA card - $699.00
  - 3 x Cremax MB235 3 bay 4 drive SATA enclosure - $225
  - 8 x Maxtor 300GB SATA drives (one hot spare) @ $262
  - Total RAID5: 6 x 300 = approx 1.8TB
  - Cost/MB: $3695/1800GB = $2.05/GB

  - Add cost of system ($1500)
  - Cost/MB: $5195/1800GB = $2.89/GB

Example 2: SATA Host-based card

4U PC with 8 SATA drives

Example 3: SCSI Host-based

- Adaptec 2200S dual channel controller @ $800

- Internal:
  - 9 x 73GB SCSI drives @ $350
  - Total: $3,950/511GB = $7.70/GB
- External:
  - 9 x 73GB SCSI drives @ $350
  - Storcase 14 bay SCSI enclosure S10A169 $5,100
  - Total: $9,050/511GB = $17.71/GB

Example of other Host-based

- Older cards still do the job:
  - Adaptec 3000 series for SCSI
  - Adaptec 2400 for up to 4 PATA drives

Example 4: SATA External

- Storcase:
  - Dual channel SCSI out
  - 12 x Maxtor 300GB SATA (one hot spare) @ $262
  - Storcase 12 bay SCSI-SATA enclosure $4500
  - Total RAID5: 10 x 300GB = approx 3TB
  - Cost/MB: $7,644 / 3000GB = $2.50/GB

Example 5: PATA External

- Nexsan:
  - Unit cost: $12,000
  - Dual channel SCSI out
  - 8 x 120GB PATA drives
  - 3 year warranty
  - Total RAID5: 6 x 120 = approx 720GB
  - Cost/MB: $12,000 / 720GB = $16.66/GB

Example 6: SCSI External - Storcase

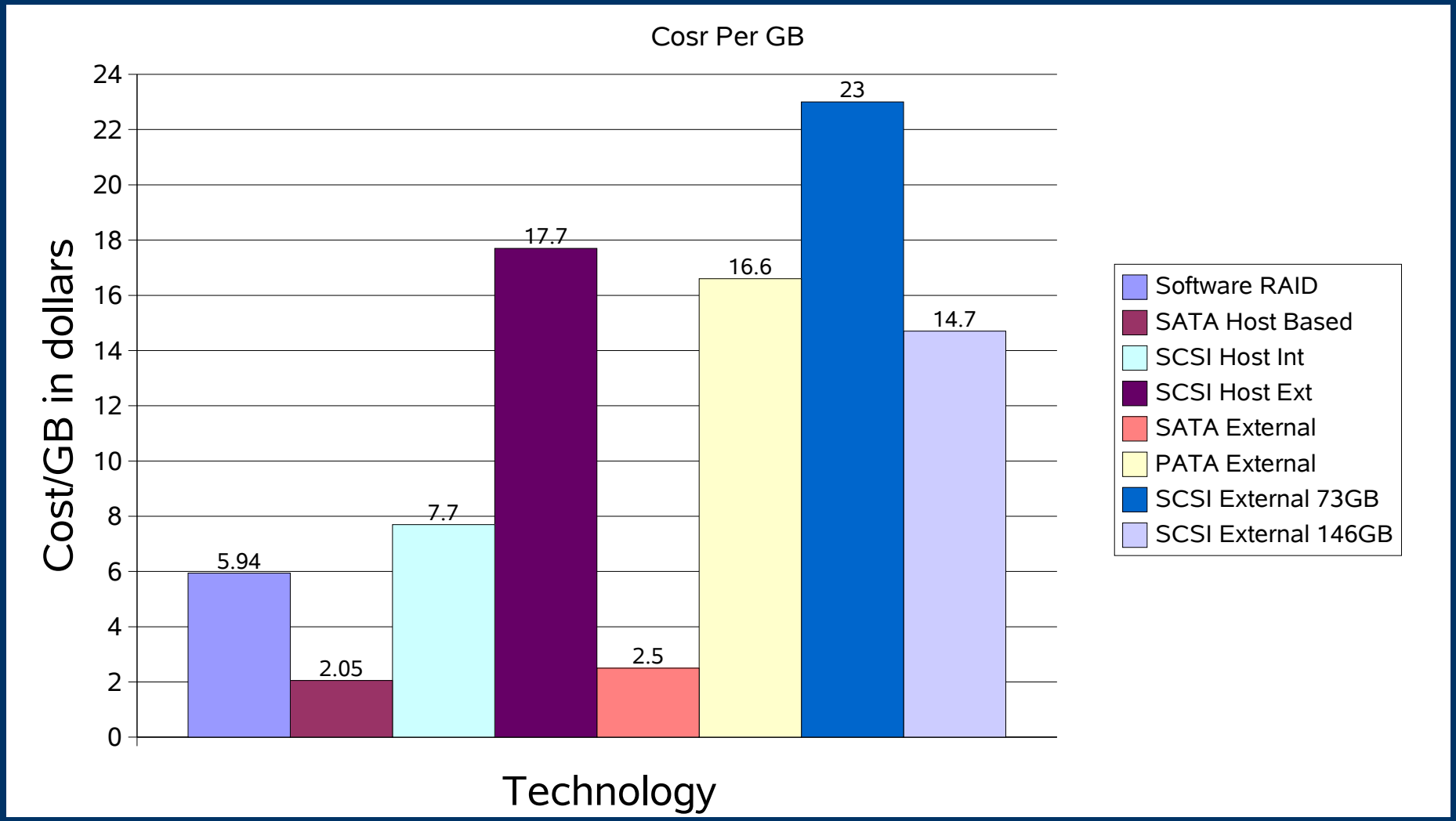
- 9 bay dual channel 320MB/sec SCSI-SCSI enclosure @ $4,000
- Single RAID controller @ $4,600

- Disks:
  - 9 x 73GB-10k @ $350 = $11,750/511GB = $23/GB

  or
  - 9 x 73GB-15k @ $638 = $14,342/511GB = $28/GB

  or
  - 9 x 146GB-10K @ $714 = $15,026/1022GB= $14.70/GB

Example Summary - Cost/GB



Some Numbers

- bonnie disk benchmark
  - Benchmark which measures the performance of Unix file system operations
  - Data collected for a wide variety of systems

- Factors:
  - CPU speed
  - Controller
  - Cables
  - Disks
  - Number of disks

Some Numbers - What's bonnie?

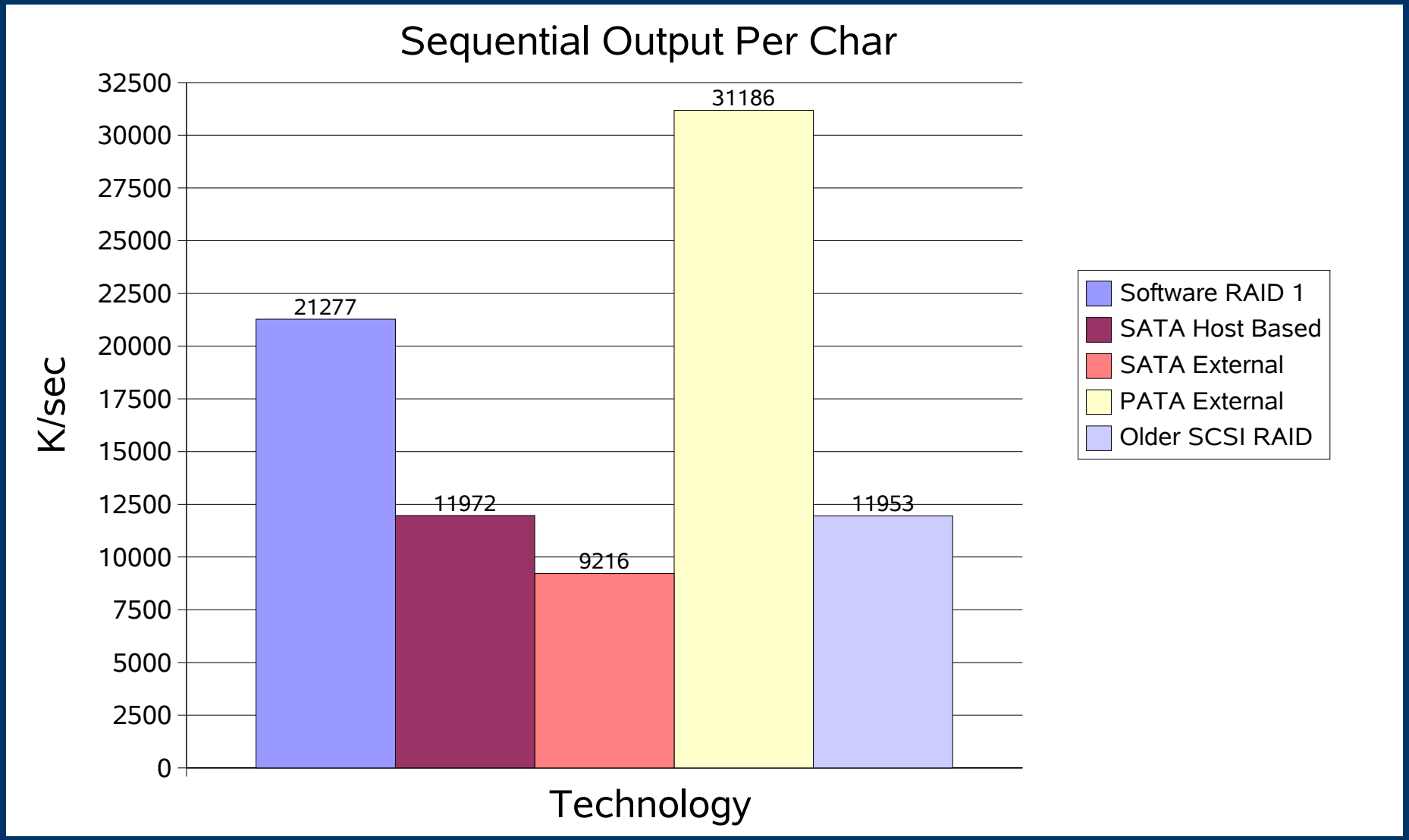
- What bonnie Does:
  - bonnie performs a series of tests on a file of known size.
  - For each test, bonnie reports the bytes processed per elapsed second, per CPU second, and the % CPU usage (user and system).
  - In each case, an attempt is made to keep optimizers from noticing it's all bogus.
  - The idea is to make sure that these are real transfers between user space and the physical disk.

Some Numbers What's bonnie?

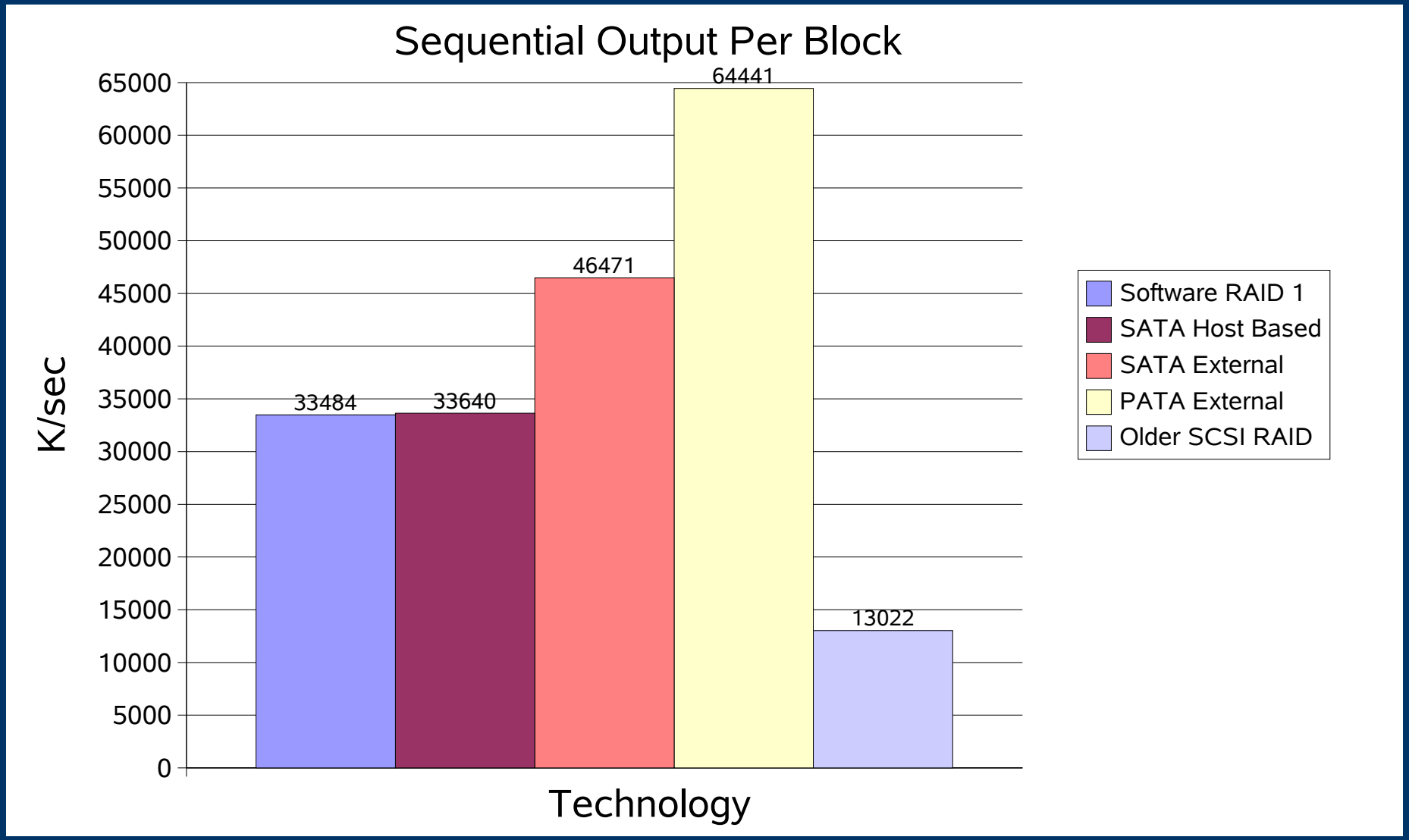
The Tests are:

- Sequential Output
  - Per-Character
  - Block
  - Rewrite
- Sequential Input
  - Per-Character
  - Block
- Random Seeks

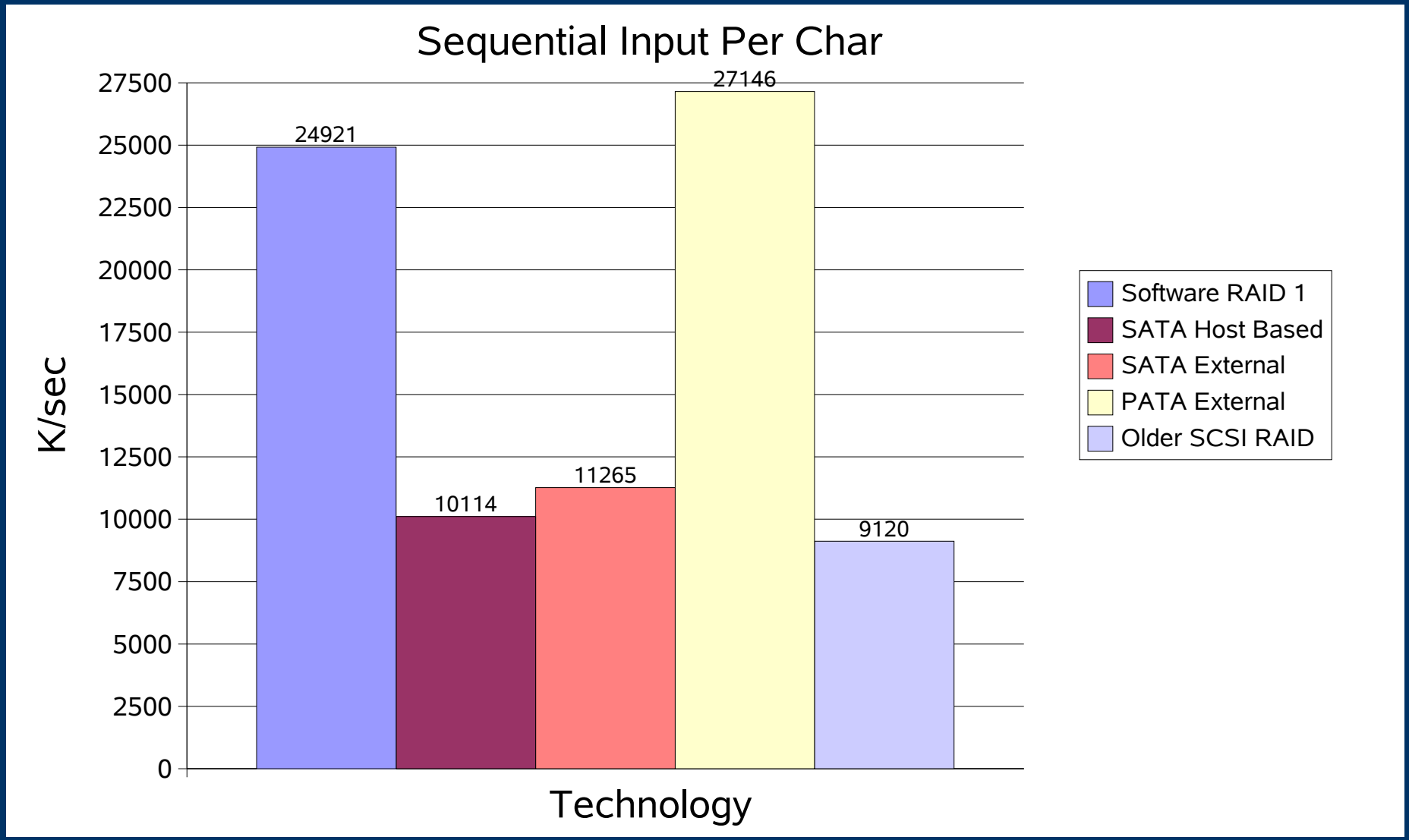
Some Numbers - Seq Output Per Char



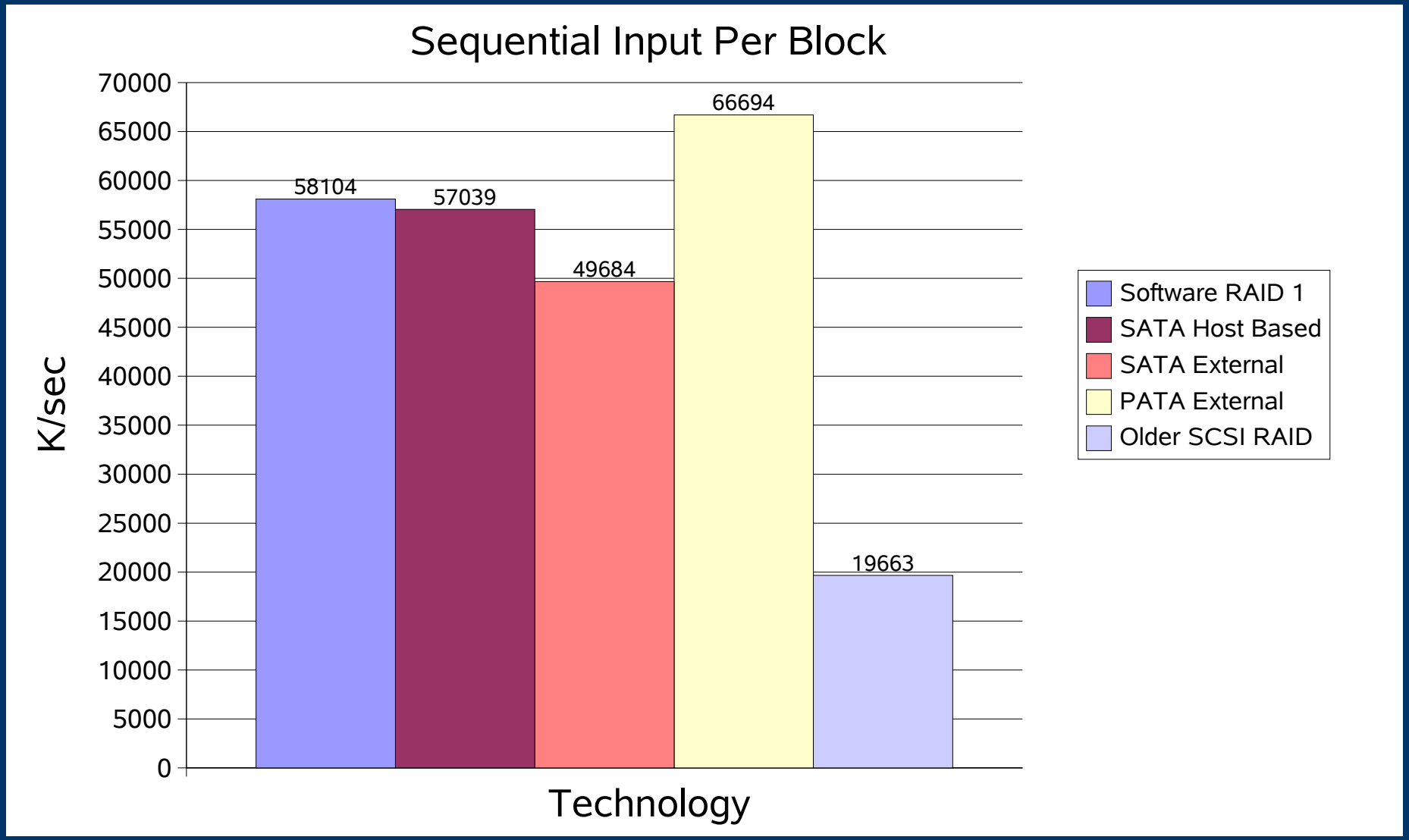
Some Numbers - Seq Output Per Blk



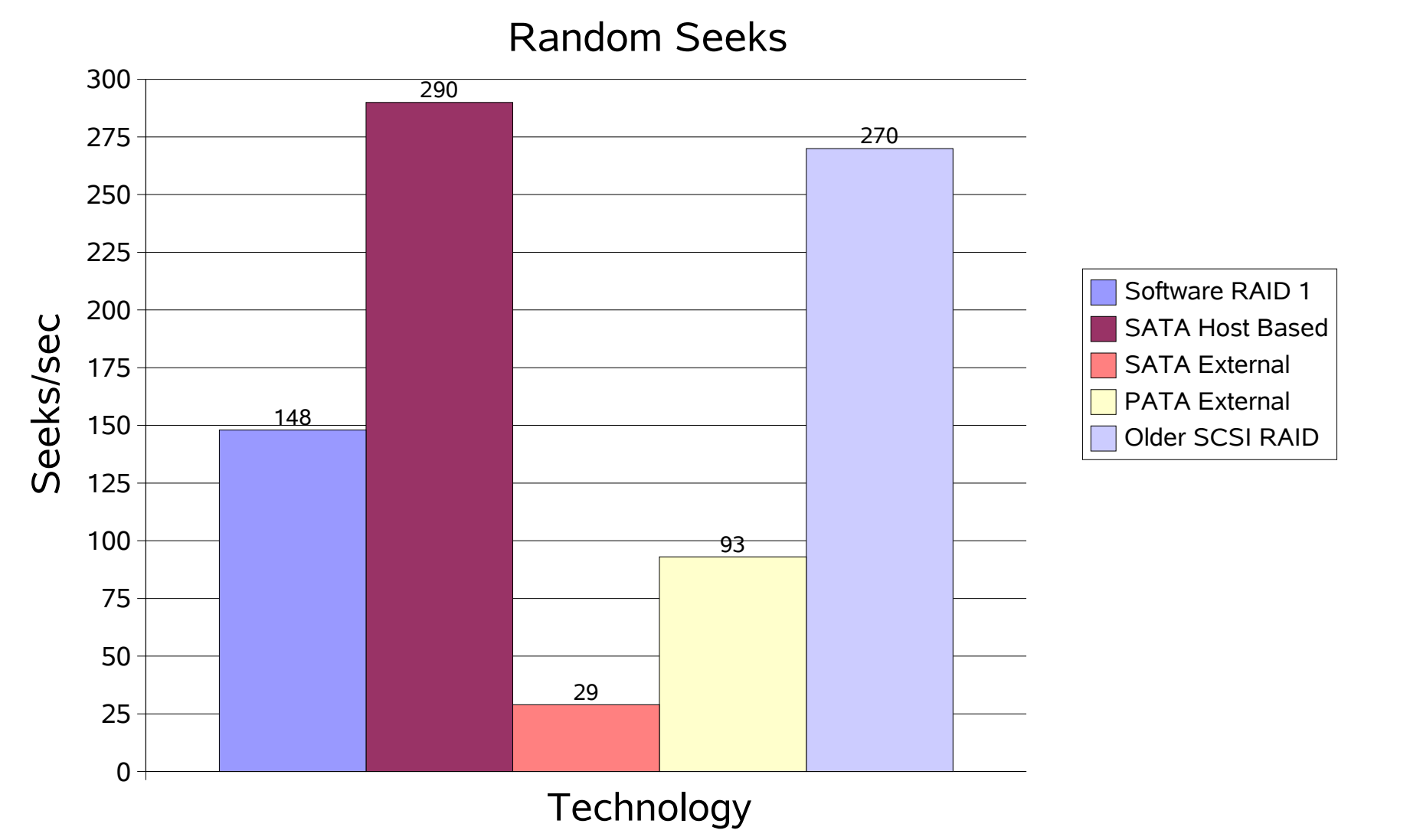
Some Numbers - Seq Input Per Char



Some Numbers - Seq Input Per Blk



Some Numbers – Random Seeks



Some Numbers - Details

```
Example 1: Software RAID
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2000 21277 76.2 33484  7.6 15322  2.6 24921 84.9 58104  4.6 148.0  0.3

Example 2: Adaptec SATA, 7 drives RAID5:
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2000 11972 99.5 33640 55.0 17119 20.1 10114 86.7 57039 27.7 290.0  4.4

Example 4: Storcase Sata 11 drives RAID5:
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU /sec %CPU
2000  9216 99.3 46471 98.5 26556 96.9 11265 99.0 49684 84.6  28.9  9.1

Example 5: Nexsan 13 drives RAID5:
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2000 31186 96.8 64441 36.3 21617 20.3 27146 91.5 66694 29.5  92.8  3.6
```

Some Numbers - Details

```
Adaptec 2400, 4 drives RAID5, Linux
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2000  8637 99.3 13724 23.3  6857  7.7  5698 68.2 15206  8.6 151.8  2.5


Single 10K SCSI drive, Linux:
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2000 25703 93.6 53043 35.6 21364 10.2 20427 82.6 85130 12.7 373.3  1.2


2 x 10K SCSI drive, RAID 1, Linux
Adaptec On-board AIC-7899 with 2005S zero channel controller
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2000 25141 91.9 53438 35.5 20786 10.0 26091 86.2 91620 22.5 479.7  3.5
```

Some Numbers - Details

```
2 x 15K SCSI disks, Software RAID0, 2 x LSI 320 bus, Solaris:
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2000 25744 71.8 119891 69.0 17707 17.0 32032 98.5 196138 80.7 357.3 5.8


2 x 15k disks, RAID1, 2 x LSI 320 bus, Solaris:
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2000 35067 96.9 63318 34.3 18055 17.5 31909 97.5 137544 53.2 345.5  4.8


2 x 15K SCSI drives, Storcase hardware RAID1, Solaris
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2047 35341 85.5 37488 20.3 17200 14.1 28931 75.3 42140 16.2  44.0  2.5
```

Some Numbers - Details

```
3 x 7200 SCSI disks, Software RAID 5, 1 bus, on Solaris:
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2000  2813 41.7  2869 17.0  1283  7.0 10693 99.4 29569 59.9  56.5 11.2


8 x Older SCSI drives, RAID5
     -------Sequential Output-------- ---Sequential Input-- --Random--
     -Per Char- --Block--- -Rewrite-- -Per Char- --Block--- --Seeks---
  MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
2000 11953 62.5 13002 10.9  5555  2.8  9120 43.3 19663  4.7 270.5  1.5
```

RAID - On a Budget

- References
  - storcase.com
  - adaptec.com
  - cremax.com
  - nexsan.com
  - textuality.com/bonnie/

RAID - On a Budget

Questions?