Tips on specifying PC hardware for Mascot

ASMS 2004

MATRIX SCIENCE

Since we are frequently asked what type of PC to purchase for Mascot, this first short session will cover some of the issues, and I'll be giving some recommendations.

Firstly, lets take a look at three possible systems

Three different PCs . . .

- 4 year old PC
  - 200s search
  - 300 Mz cpu
  - Dual CPU
  - 256Mb RAM
  - $49 on ebay

- IBM 'A' series
  - 15s search
  - 3 GHz cpu
  - Single CPU
  - 2 Gb RAM
  - $1,600

- IntelliStation Z
  - 14s search
  - 3.2 GHz cpu
  - Single CPU
  - 2 Gb RAM
  - $4,600

ASMS 2004

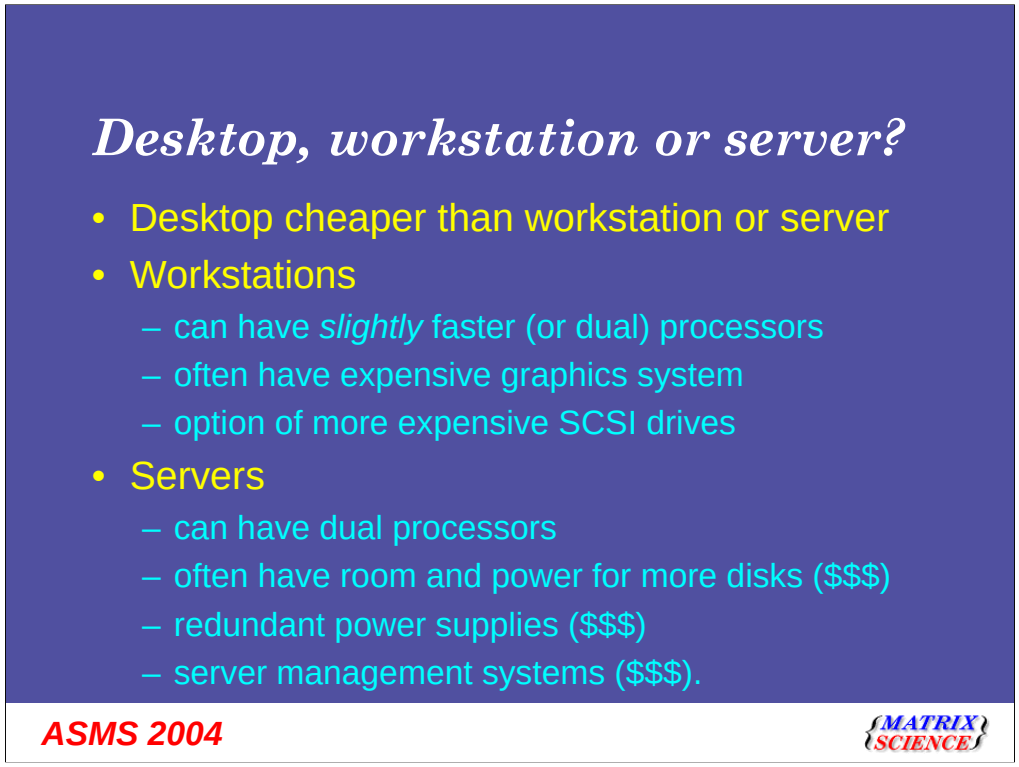
MATRIX SCIENCE

Sadly, quite a few of our customers spend a not insignificant amount of money on Mascot licences, but then load the software onto an old spare PC lying around in the lab. For example, a typical search on a four year old system with a 300MHz processors, a fancy disk array but only 256Mb RAM might take 400 seconds. Although this system may be one of your favourites, it would probably only sell for about $49 on EBAY

Next on the block is a budget system from IBM - our search time is down to 15 seconds, and the system cost quite a modest $1600

And finally, a top of the range single CPU system from IBM costing $4,600 - but the search time has only increased by 1 second.

I hope you are already convinced that it's worth spending sixteen hundred dollars - but when is it worth spending the extra 3k?



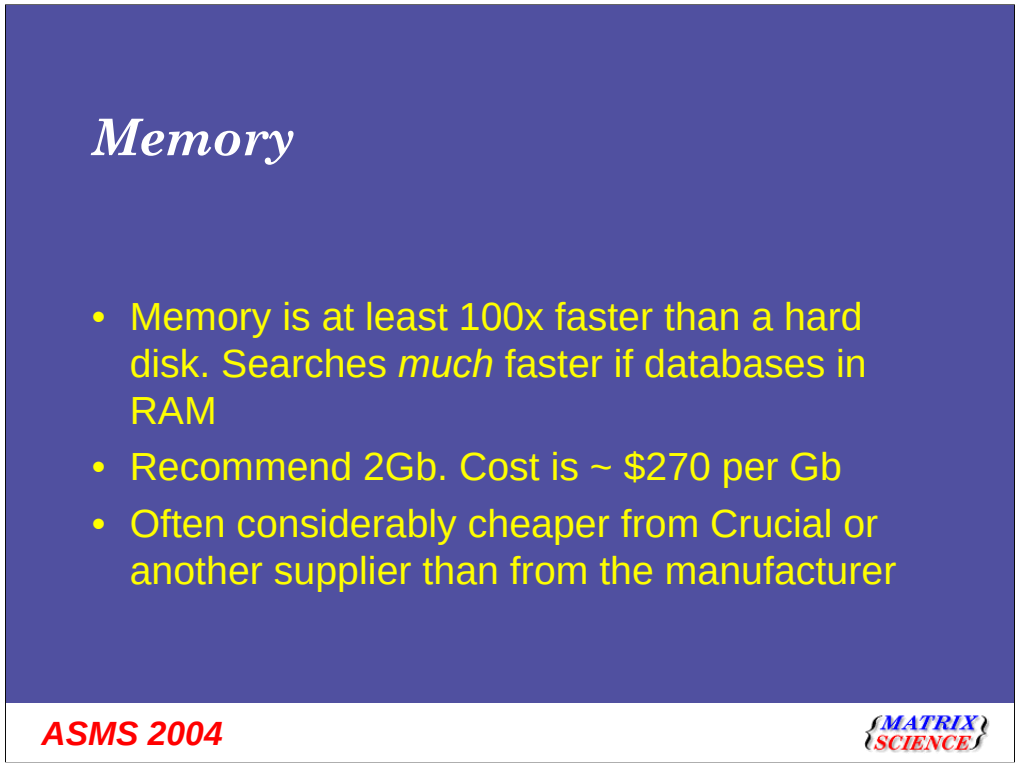
If you are looking on the IBM or DELL web sitefor a PC, the first choice you have to make is whether you want a desktop, a workstation or a server. These terms are primarily "marketing" terms, but there are some general differences.

Desktops are mostly cheaper and smaller than workstations.

Workstations are often touted as top of the range desktops - indeed they sometimes, but not always have slightly faster processors. In some cases, there will be a dual processor option that won't be available on a desktop. Most importantly, they will have a classy graphics system - which is of course totally useless for Mascot and most mass spec software. They will also often have SCSI drives, that are also not really required for Mascot. So, we can see that if you are only looking for a single processor system, it's best to avoid the work-stations

Servers can also have dual processors. They tend to be larger with plenty of room and power for mutiple hard disks. They may also have two or more power supplies in case one fails - called redundant power supplies. Finally, they often have server management software that will only be of real use to you if you have the time and skills to learn how to use it. The server management software will not make Mascot run faster.

In summary, it is normally best to choose a desktop if you want a single CPU system. If you want a dual processor system, then you will need to look through the server and workstation options for the best system

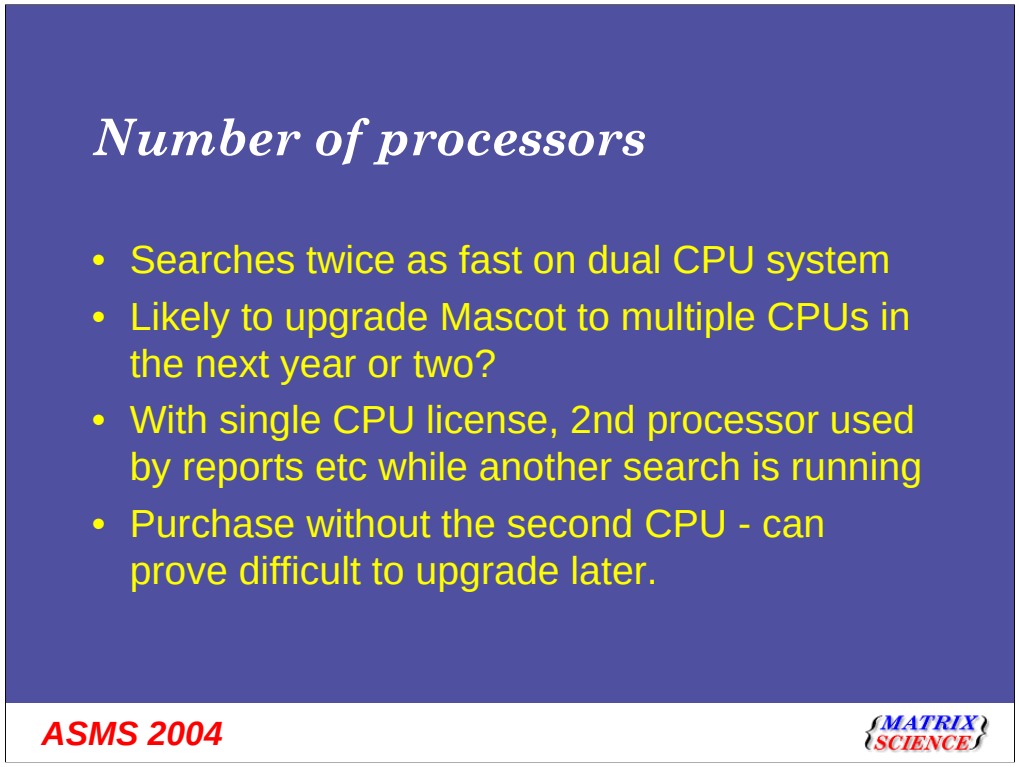


In a Mascot search against the NCBI non-redundant database, the search engine needs to read through 600Mb of sequence data. On a fairly recent IDE disk, it takes about 25 seconds to stream through this data. It takes a fraction of a second to stream through this in memory, so each search would take nearly 25 seconds longer from disk than if the database was held in memory.

A far worse situation is if two searches are running at the same time, because the disk heads need to keep jumping from one spot to another. In this case, it takes an amazing 7 minutes to stream through the database - quite a long time to wait for a peptide mass fingerprint that should maybe take 15 seconds.

We recommend 2Gb RAM - this is really a small cost compared with the cost of the instrument or the Mascot software. It is now possible to put more than 2Gb in a PC, and Mascot will use this if it is set up properly.

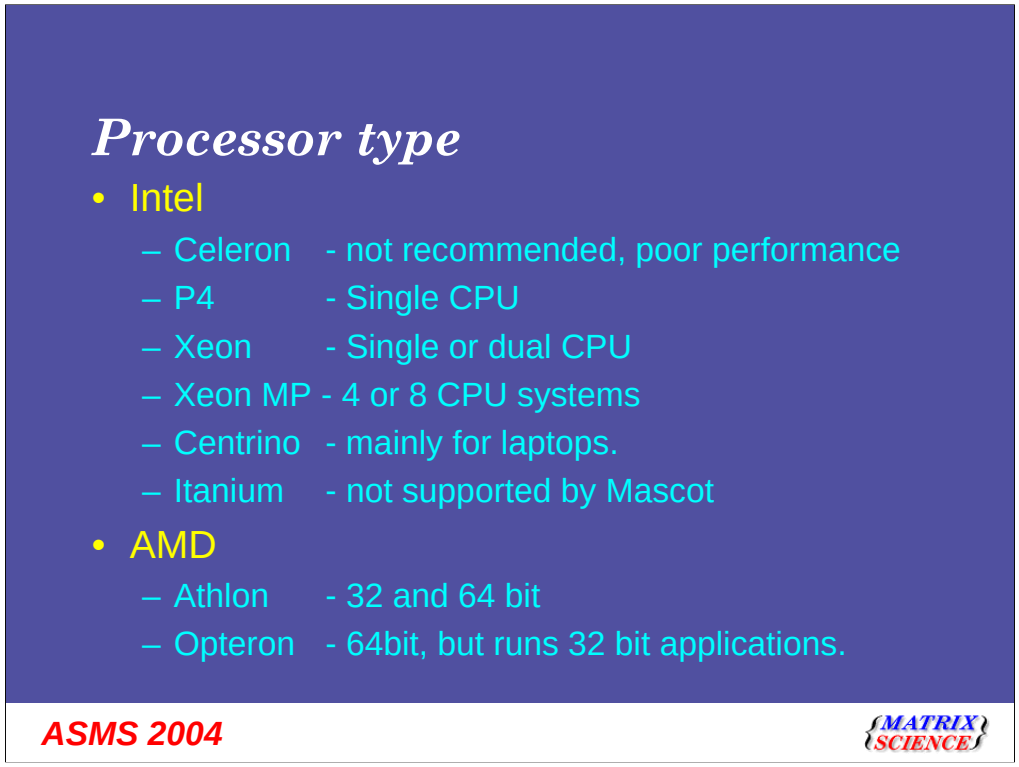
Memory is often more expensive from the computer manufacturers - can be half the price if you are prepared to install it yourself.



A Mascot search will run nearly twice as fast on a dual CPU system. If you have a dual CPU Mascot license, then you obviously need a dual CPU system. If you have or are intending to purchase a single CPU system, is it worth buying a dual CPU server?

The answer is yes if there is a possibility of upgrading to a 2 cpu license in the next year or two or if you do large searches that can produce very large search results, or if several people are likely to be using Mascot at the same time.

Finally, you can purchase a system with space for a second CPU and intend to buy the second CPU later. While this will probably be OK, you should be warned that it may be hard to buy the second processor of the required speed two years later.



There are two main processor manufacturers - Intel and AMD. Intel currently produce 5 broad types of processor that you may find in a PC:

- Celeron. You'll find this in the budget desktop systems. We have never tested Mascot on this processor, but would expect poor performance because of the small cache size.

- Pentium 4. The standard processor for a single CPU system. Currently available in processor speeds of up to 3.4 GHz. Some of these processors have support for Hyper-Threading technology that I'll discuss in a moment.

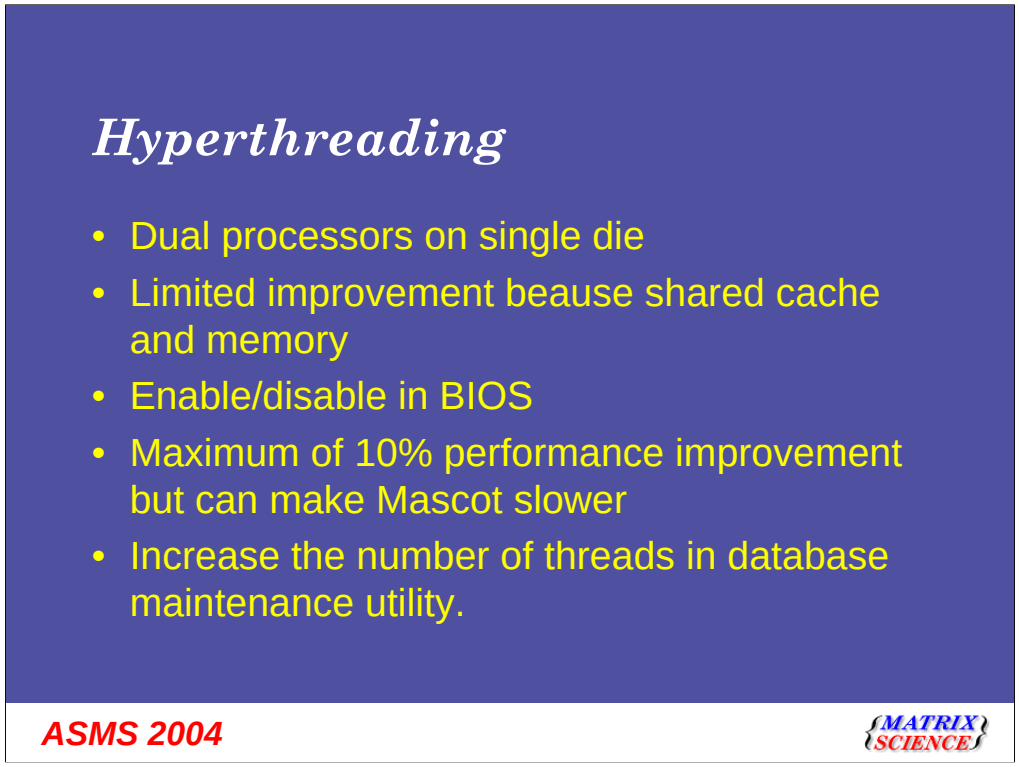
- Xeon - speeds up to 3.2 GHz, and with varying cache size. Support for up to two processors

- Xeon MP - speeds up to 3.0 GHz. Support for 8 CPU or more systems. Tend to be very expensive, and are slower. Not recommended.

Centrino systems are low power and mainly in laptops - and hence are not currently relevant

- Itanium is a 64 bit processor. Unfortunately, it's rather slow for integer calculations required by Mascot and we therefore isn't supported.

AMD currently produce 2 processor families - the Athlon processors and the Opteron processor. There are 64 bit and 32 bit Athlon processors, while Opteron is just 64 bit. We don't currently have a version of Mascot to run in 64 bit mode for these processors, but it does run in 32 bit mode. We don't yet have performance figures for the Opteron - but feedback from customers indicates that the performance should be at least comparable, if not significantly better than with Intel processors.



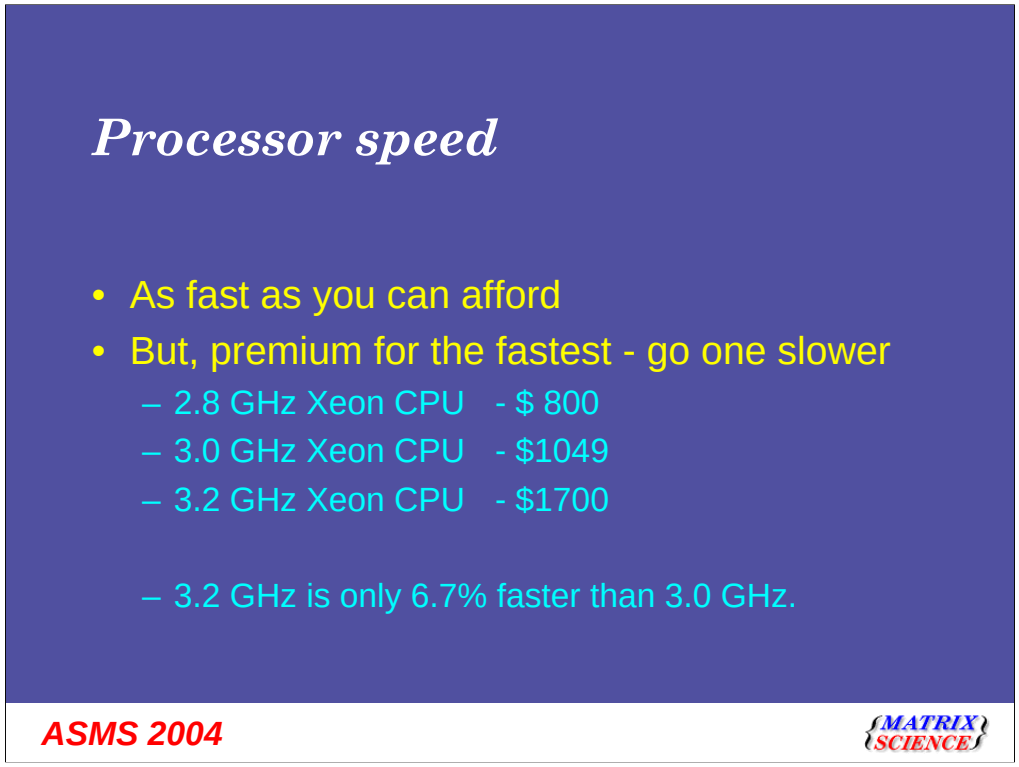
Recent Pentium 4 processors actually have two CPUs on one processor chip. Intel have called this hyper-threading. You may think that having two CPUs instead of one would give you twice the performance - however, there is limited improvement because the memory and some other resources are shared. Intel now claim that you can see a performance increase of up to 25% - but most people report much lower numbers than this.

Hyper-threading is normally enabled/disabled in the bios - you will need to refer to your computer manual for details.

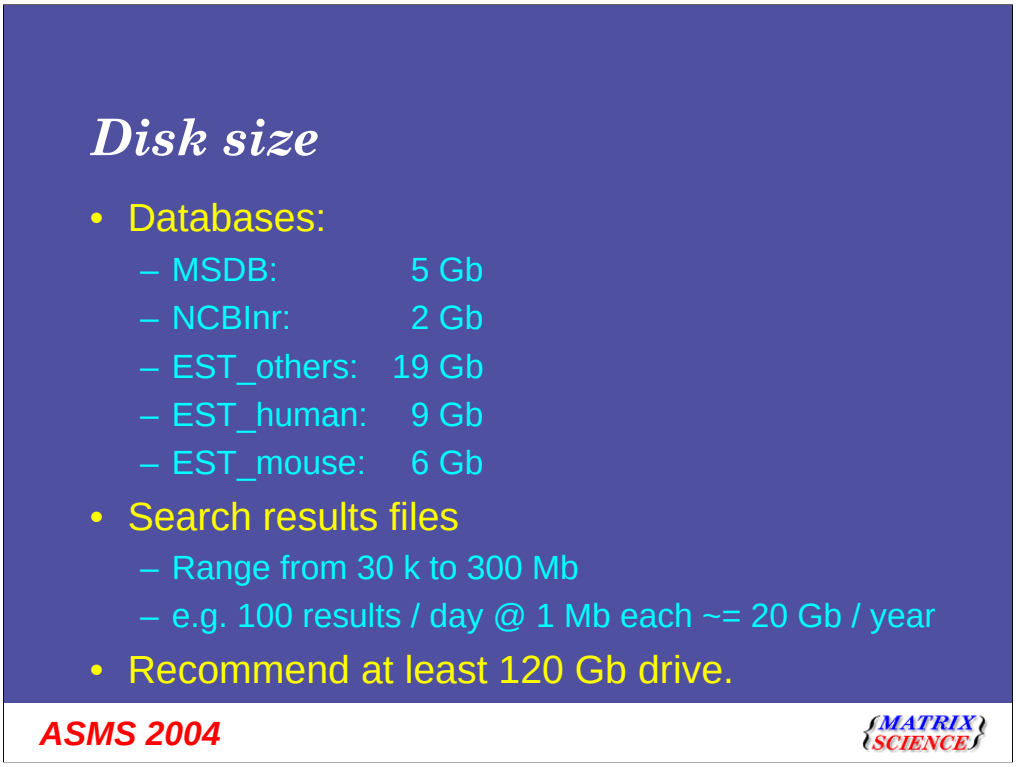
If hyper-threading is enabled, the Mascot status screen (version 2.0 and later) should show this.

The best performance improvement that we have seen with ht is 9.8% - however, in some situations, Mascot will actually run slower with hyperthreading enabled. For Windows, you seem to need to be running XP or Windows 2003 to get this improvement. For Linux, make sure you have a kernel 2.4.19 or later

If you have a single CPU license, and enable hyperthreading, you will need to set the number of threads to 2 in the database maintenance utility.



The basic rule for processor speed is buy as fast as you can afford. However, there is always a premium price for the very fastest CPU - look at the prices here and see how it jumps for the 3.2 GHz processor which is only 6.7% faster than the 3.0 processor.

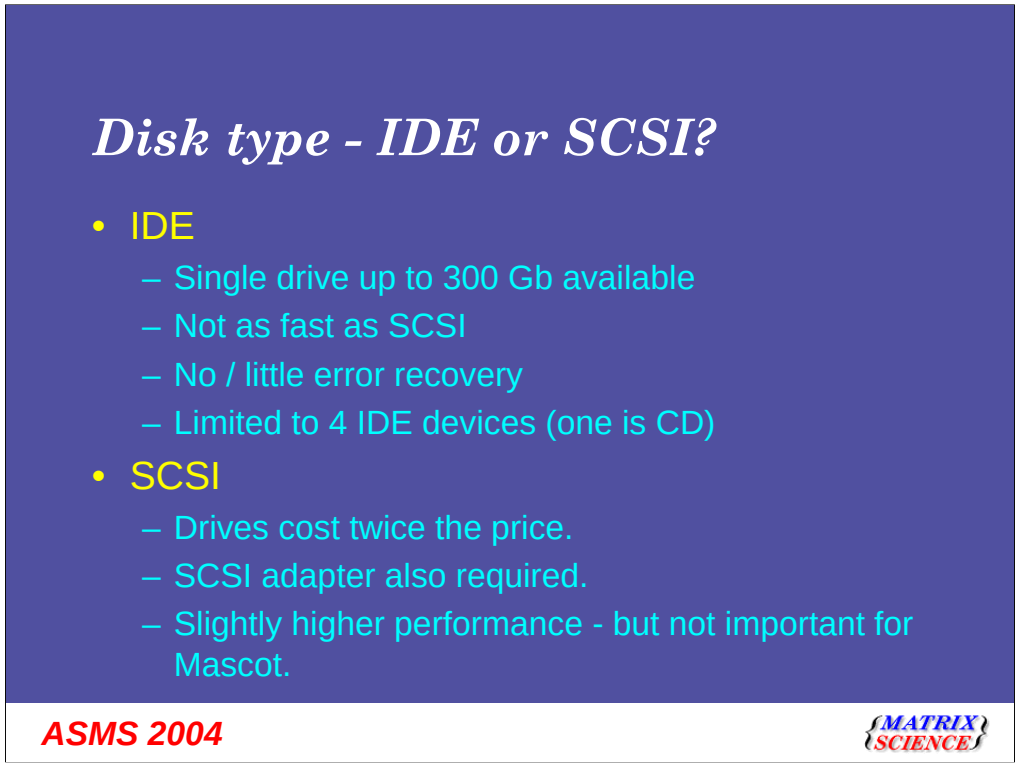


The Mascot server disk is essentially going to be filled up with fasta databases and Mascot results files.

Database sizes are still increasing rapidly - as you can see from the list above, if you have MSDB, NCBInr and the three EST databases, then you will need a total of 41 Gb. These sizes include the size of the compressed files, and one 'backup' in the old directory.

The size taken up by search results files is not so easy to calculate - it all depends on how many and what type of seaches you perform.But, for example, if you save 100 results files per day, and they are each about 1Mb, then this adds up to 20 Gb per year.

Based on the above, we would recommend at least 120Gb



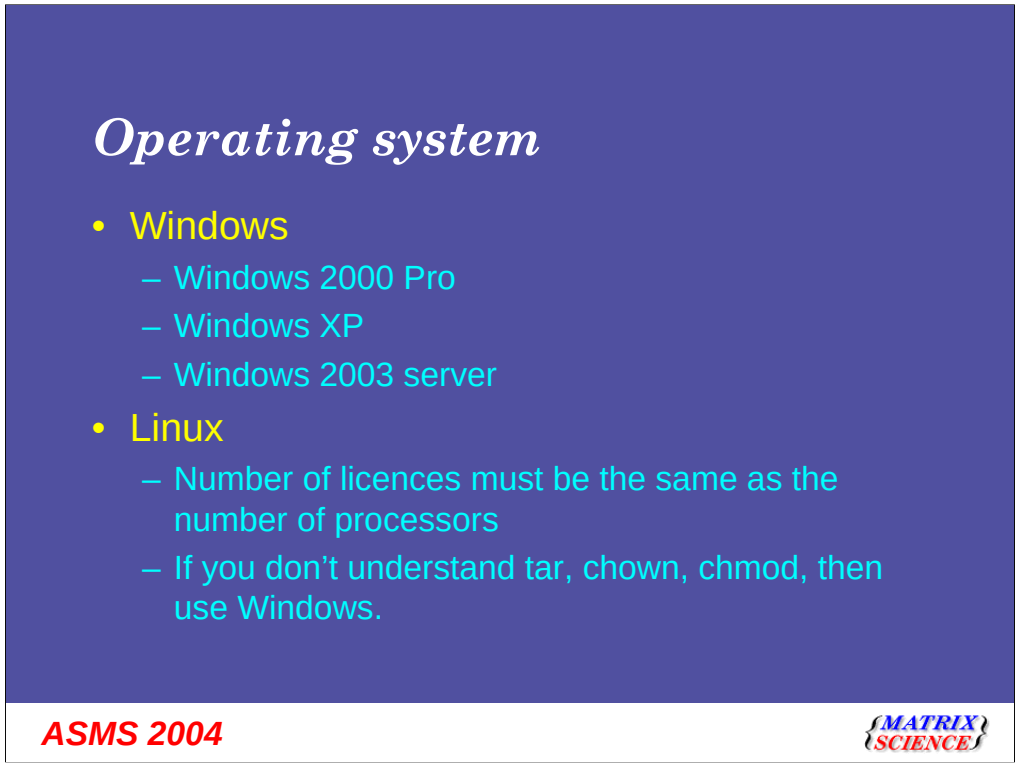
There are two types of disk - IDE and the more expensive SCSI

180Gb IDE drives are now quite common, and 300Gb drives are available (although they are a little slow). This disadvantages over SCSI is that they are not as fast, and there is little or no error recovery available with them. With a single IDE controller, there is a limit of 4 IDE devices, and since one will be the CD drive, it is best to choose large drives.

So, SCSI is at least twice the price of IDE

However, as we have seen earlier, if you have enough memory in the computer, then you don't need fast disks. If you have a SCSI RAID system, then it's possible to set it up so that if a disk fails, then no data will be lost. This obviously also adds to the cost.

In general, it's not worth paying the extra for SCSI

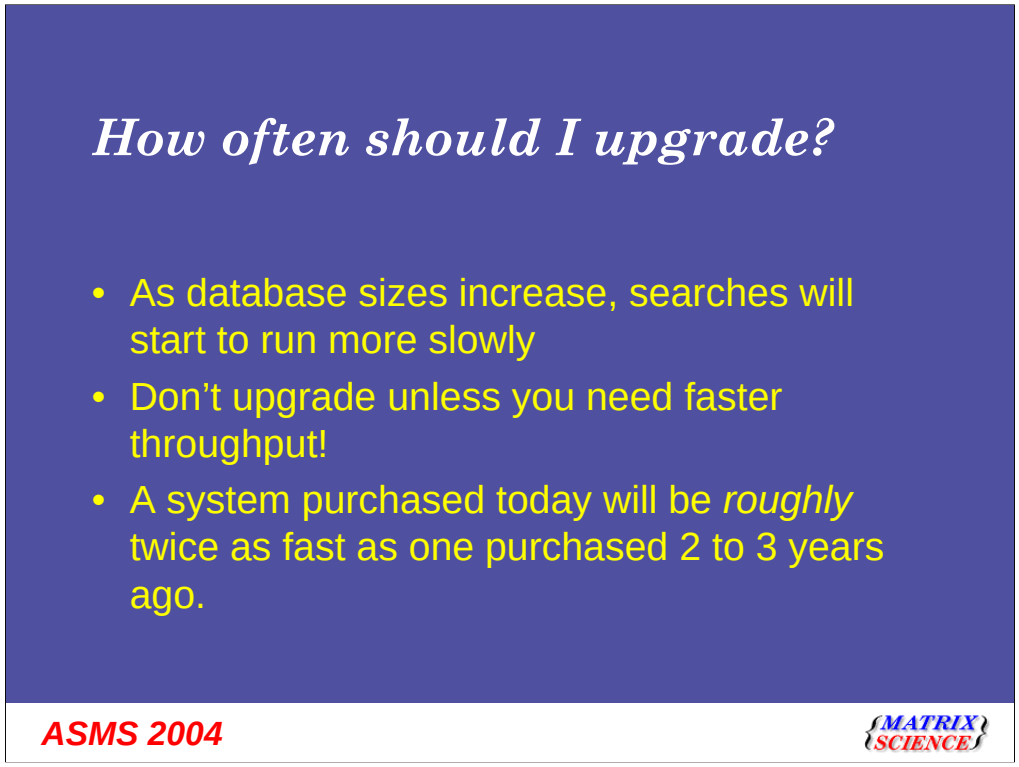


For standard PCs, the choice of operating system is Windows or Linux. For Windows, choose 2000 PRO, XP or 2003 server. We don't recommend Windows 2000 server if you have a dual CPU system - there are some scheduling issues that can cause poor performance which have been resolved in Windows 2003 server.

You only need to choose the server options if you want more than 5 simultaneous connections to the web server.

For Linux, you should note that the number of Mascot licences must equal the number of processors in the system.

We are often asked whether we recommend Windows or Linux. The performance is identical on both platforms. However, if you haven't used Linux before, then we strongly recommend using Windows.



Over time, searches will run more slowly because the database sizes are increasing.

In Feb 2002, MSDB had 800,000 sequences.

Two years later, it had nearly double that number of sequences - however, a system purchased today will be roughly twice as fast as one purchased in Feb 2002.

For many people, upgrading their system about every two to three years is sensible and cost effective.

Conclusions

- For a single CPU system:
  - 'Desktop' will be cheaper
  - 3.x GHz processor, 2GB memory, 120Gb IDE disk
- For a dual CPU system:
  - Need a 'workstation' or 'server'
  - 3.x GHz processor, 2GB memory, 120Gb IDE disk
- Aim to replace the hardware every 2 to 3 years.