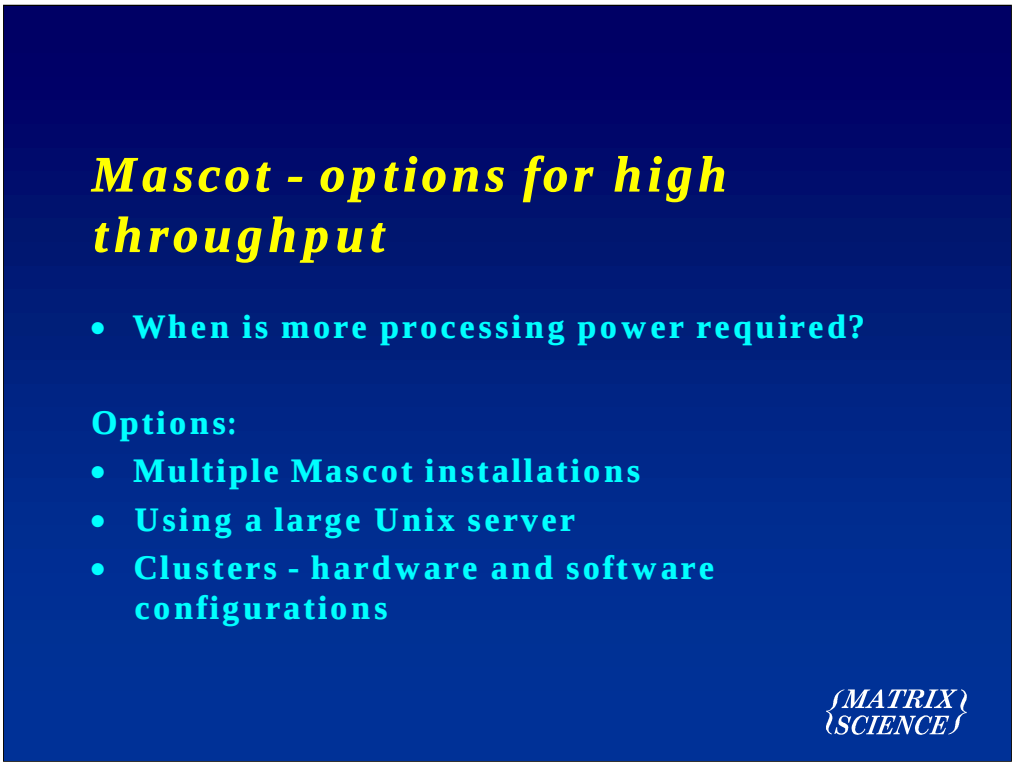
Scaling from 1 PC to a super computer using Mascot

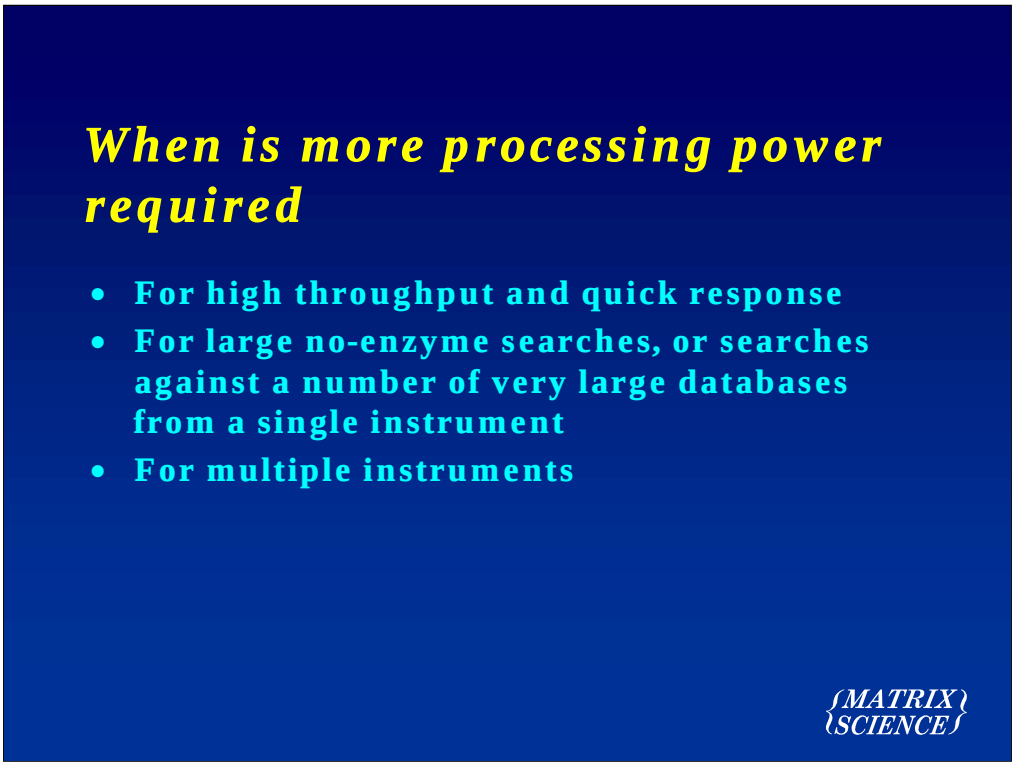
MATRIX SCIENCE



In this session I am going to talk about the three main options available if you need higher throughput for Mascot searches than is available on a standard PC.

So firstly I will describe how to determine if you need more processing power and if so how much. Then I will discuss the three options:

- Multiple installations, which we don't generally recommend
- using a large Unix server
- using a cluster

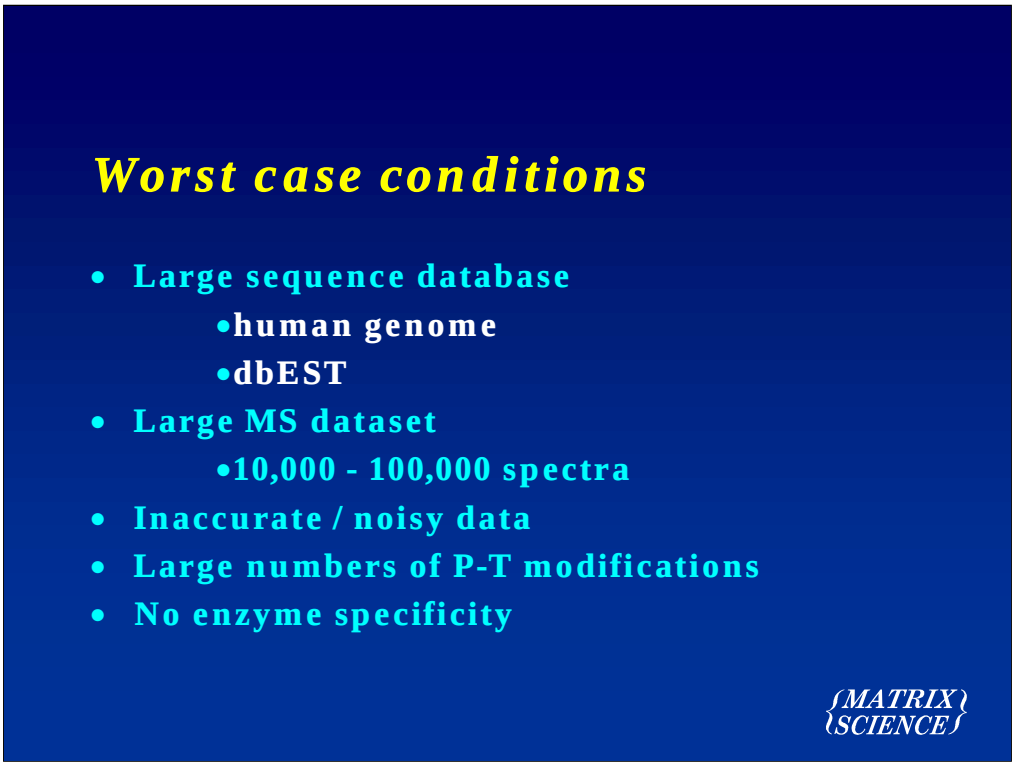


So firstly, why might you need more processing power. You may think that this is all very obvious, but I can assure you that we get a fairly equal number of requests for quotes for overpowered and underpowered systems.

The main reasons are listed here:

Clearly if you have over 50 instruments running in a high throughput environment like ...

GeneProt, then a single PC running Windows is not going to be up to the job.



So what are the worst case conditions -

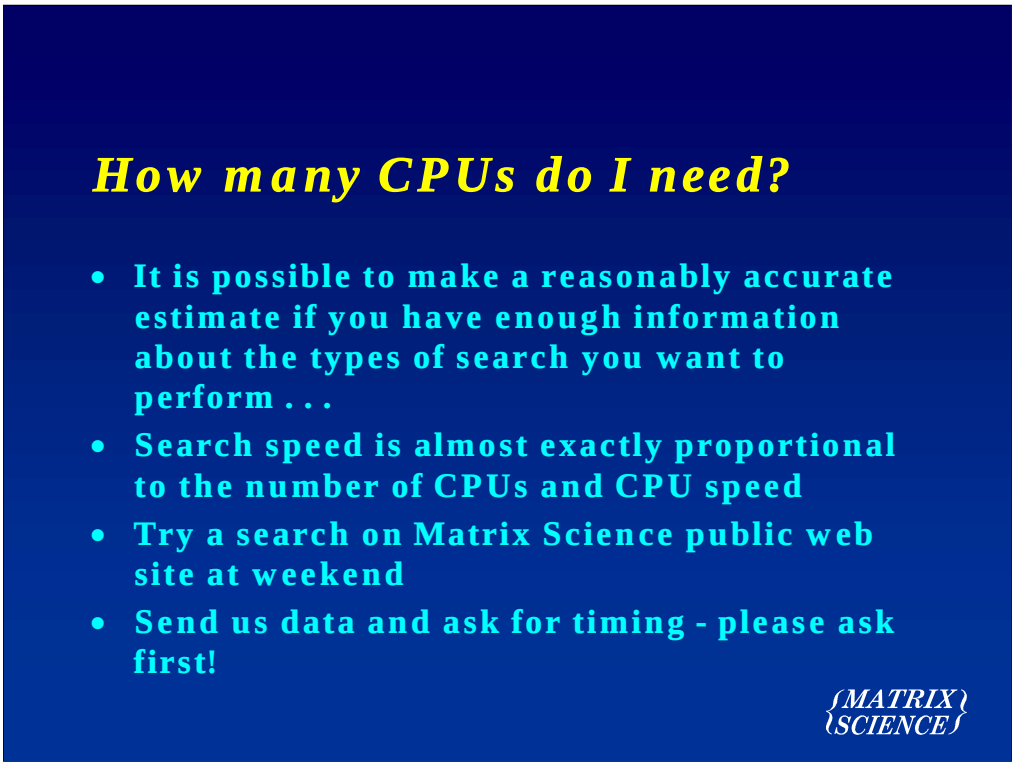
The human genome has approximately 3 Gigabases, and the complete dbEST is now over twice that size. Searching these databases is clearly going to be extremely computationally intensive.

However, even a single instrument running a MUDPIT experiment can produce in excess of 10,000 spectra.

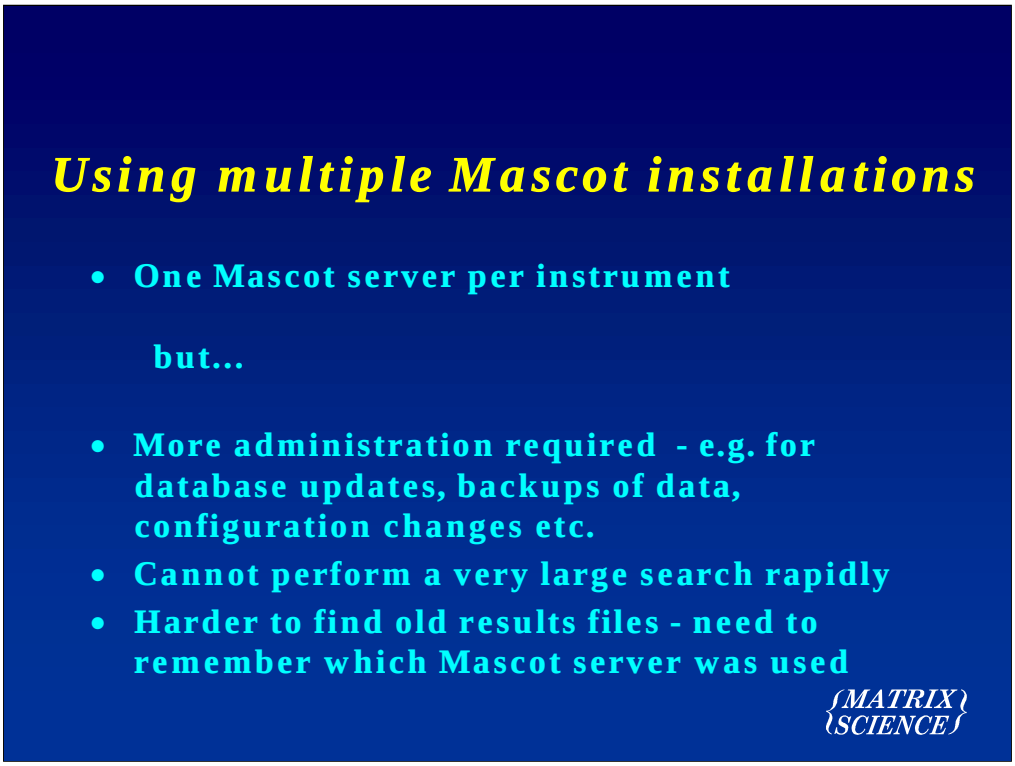
Inaccurate or noisy data can mean using large tolerances, and this will affect the search speed.

Large numbers of post translational or chemical modifications (selected as variable modification) will also increase the search speed. There is no time penalty for fixed modifications.

Lastly, a search with no enzyme specificity will dramatically increase search times.

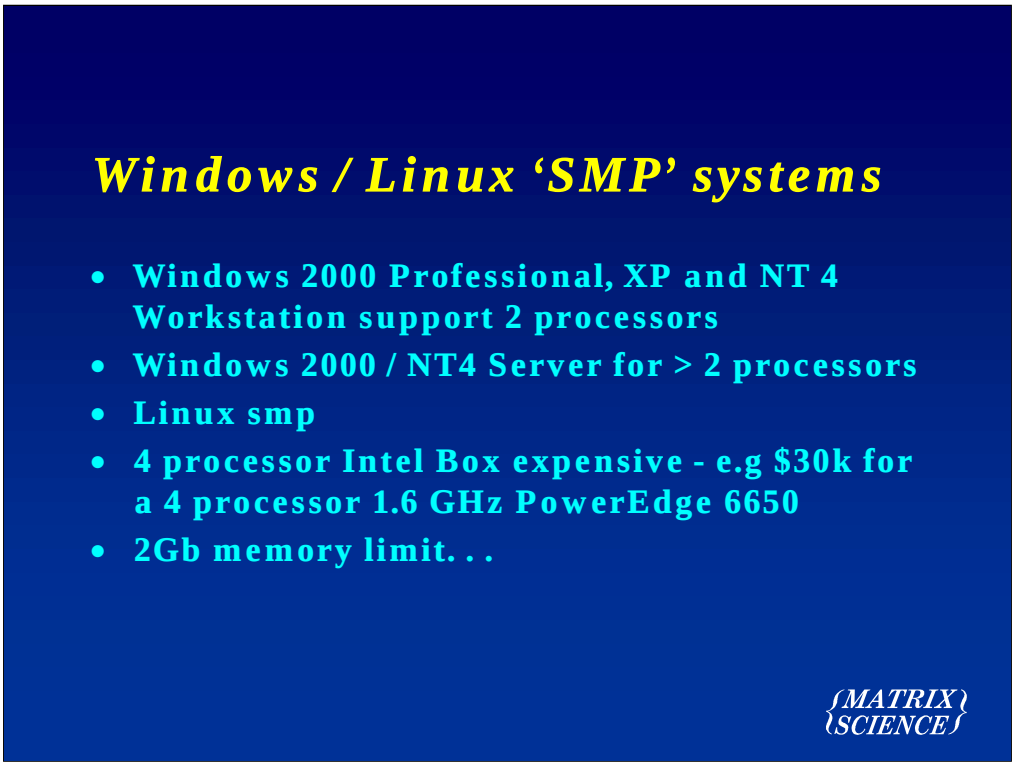


Estimating the number of CPU's required to achieve a given throughput is complex. One option is to try typical searches on our web site.



So the first, and most obvious solution is to have multiple Mascot installations - for example one per instrument, or one per user.

However, we would generally recommend against this because...



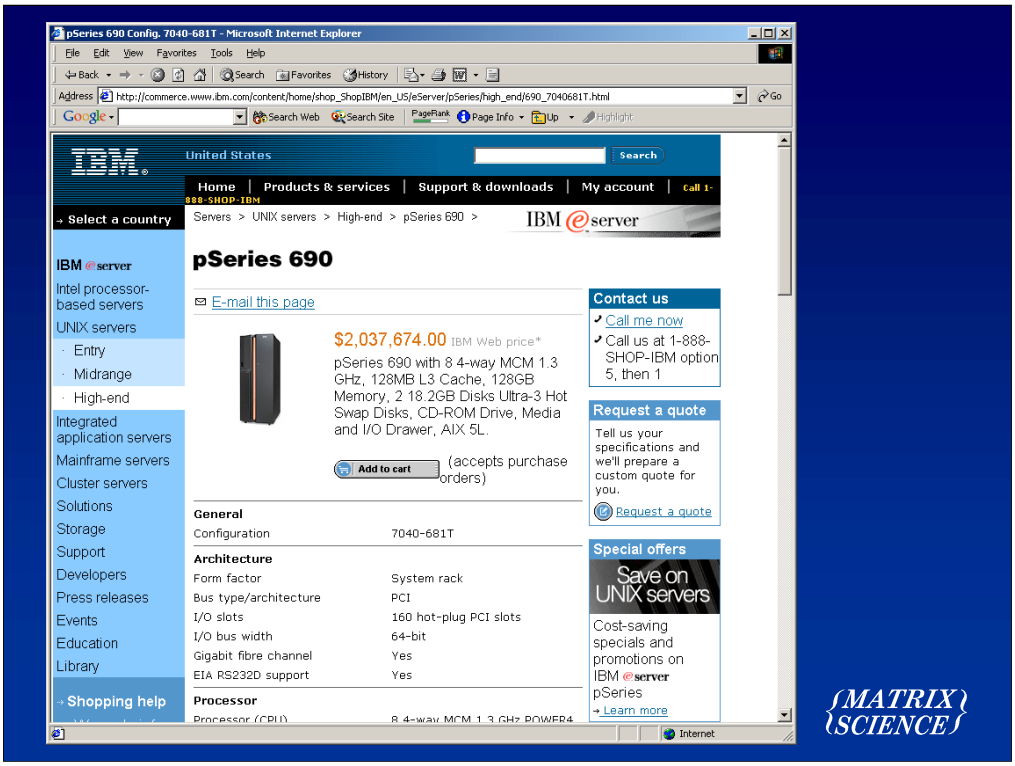
For a simple Intel or Athlon based PC, the maximum number of processors that we would recommend is 4.

With Windows 2000 professional, XP and NT4, the limit is 2 processors.

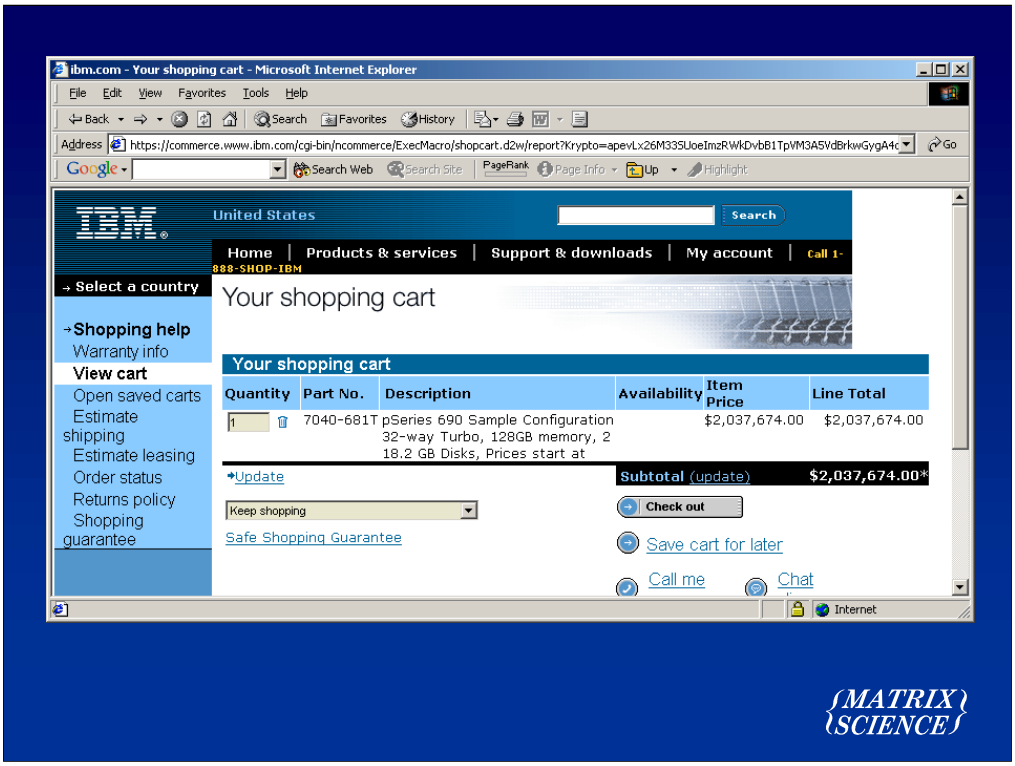
So then we have the big Unix Box solution.The four main manufacturers are shown here, and each system has its merits.

The advantage of these systems is that they really can scale.

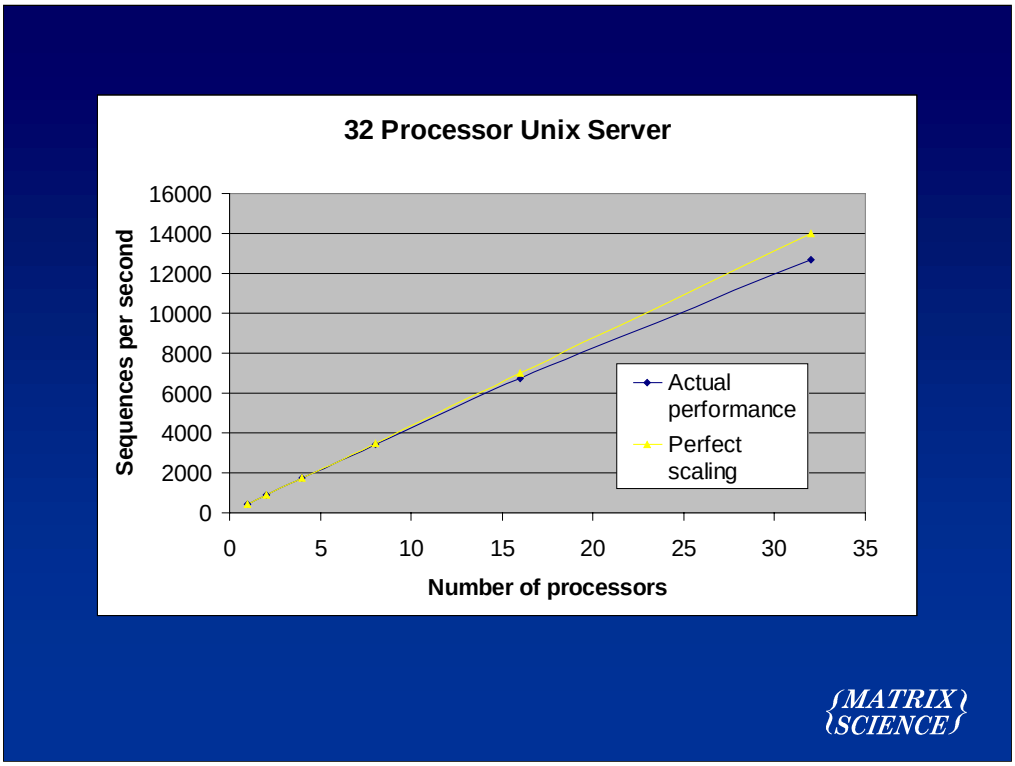
. . .



And what's more, they can now easily be purchased over the internet



Anybody got a credit card I can borrow!



This graph shows how well one of these systems can scale. This graph is for a 32 processors Unix box, and the yellow line is showing perfect scaling. Once we get up to about 32 processors, it starts to tail off a little.

Advantages of a large Unix system

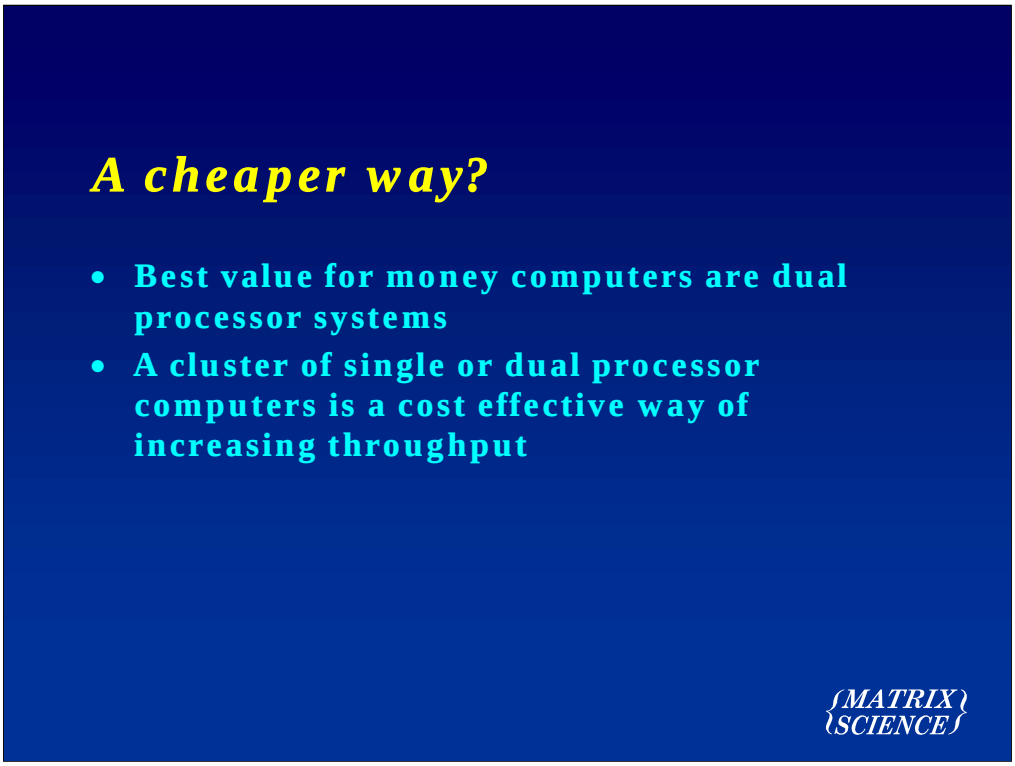
- Often ‘company’ fasta files compiled on a large Unix box
- Robust / excellent support / well maintained
- Some systems can scale to more processors
- In house support from IT / Bioinformatics
- Large amount of memory
- Easier to maintain and support than a cluster

MATRIX SCIENCE

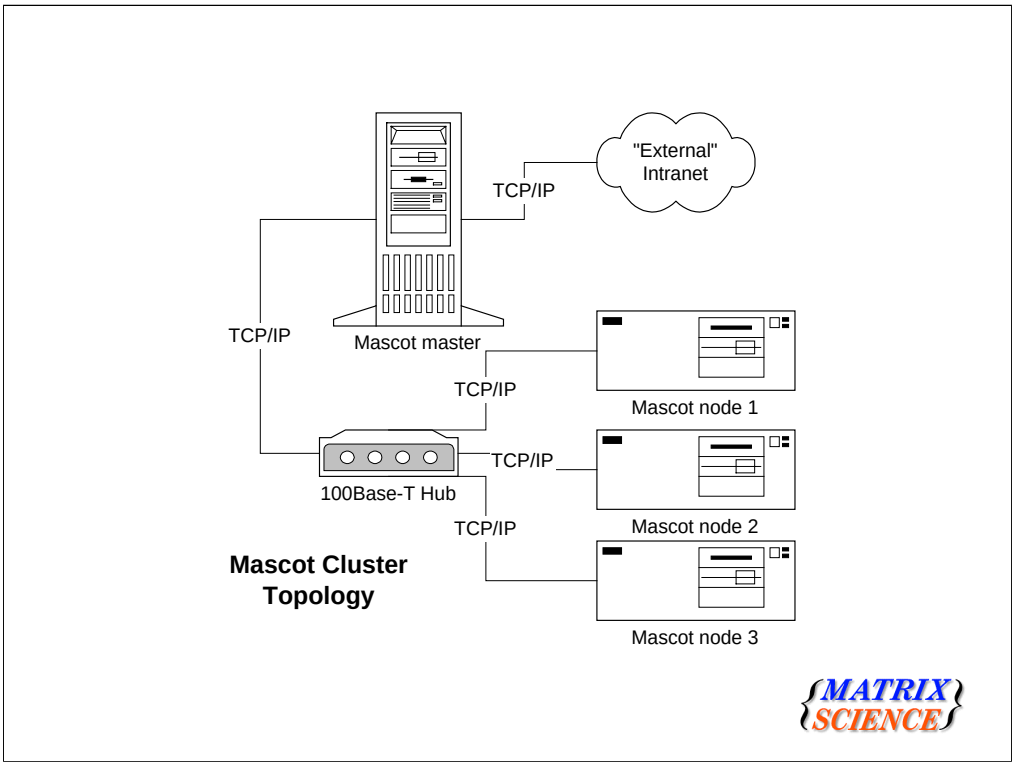
Disadvantages of a large Unix box

- **Older systems can be slow - 2 year lifetime**
- **Price**

MATRIX SCIENCE



A cluster of standard PC's is a very cost effective and scalable solution

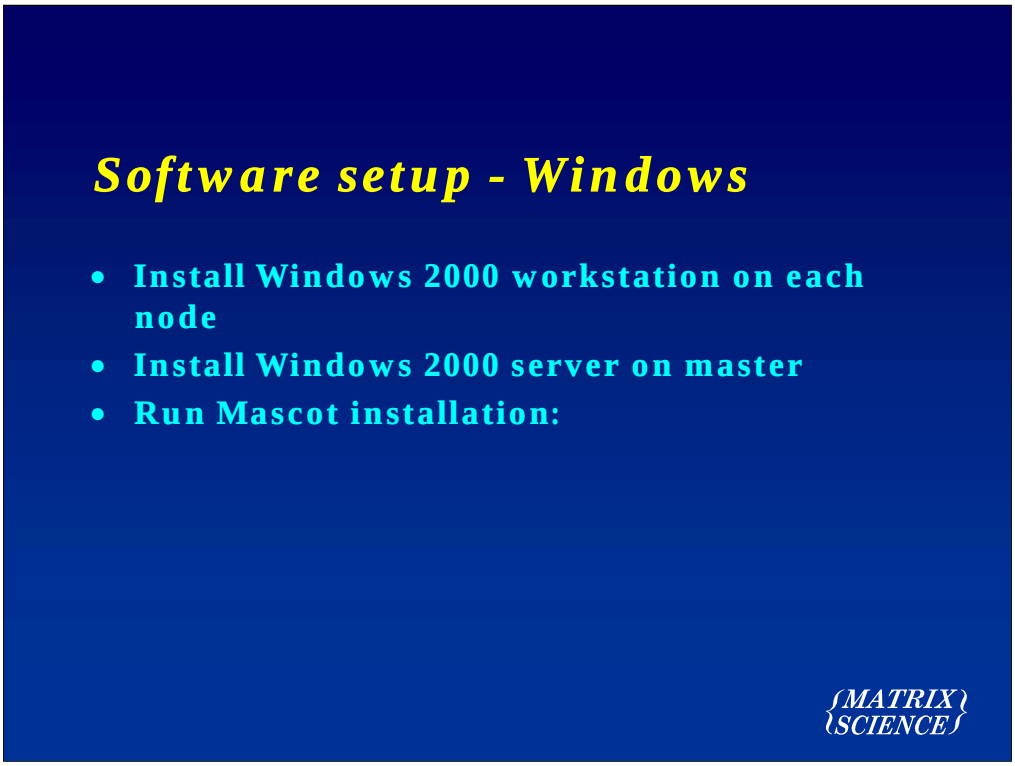


Mascot supports cluster operation using a Beowulf-like topology. Mascot supports cluster mode as standard, whenever the licence is for 4 CPU's or more. You just have to hook up an appropriate number of PC's on a local LAN

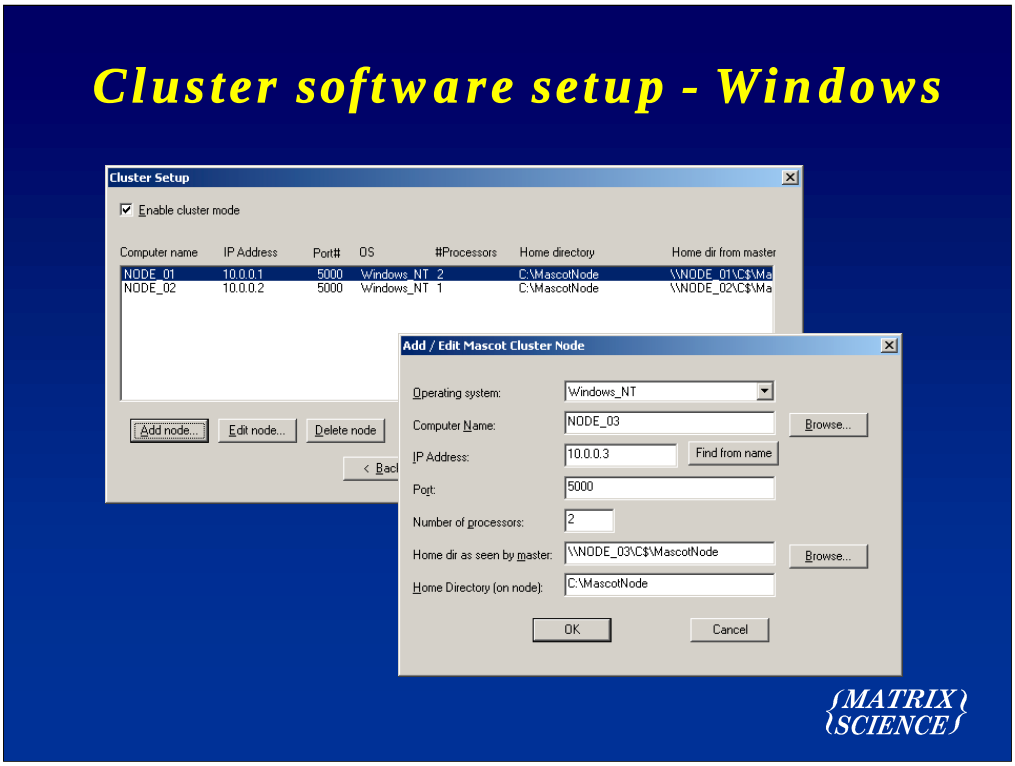
Matrix Science Web server

- **19” rack-mountable chassis containing 10 fast Intel processors**
- **Redundant power supplies**
- **We run Windows on the nodes, and the ‘Master computer’ runs FreeBSD**
- **2Gb RAM per node**
- **Hot-swappable**
- **Similar modules available for sale as a turn-key package**

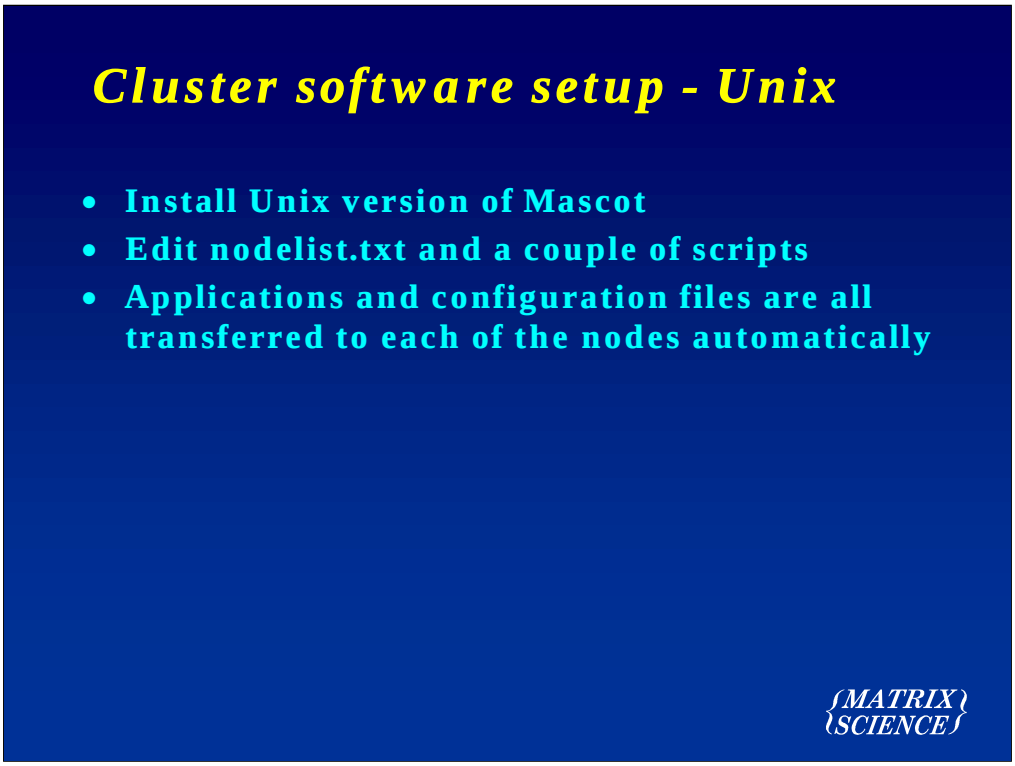
Alternatively, we can supply a complete turn-key solution based on rack mounted, high-reliability PC hardware. Our web server uses this hardware



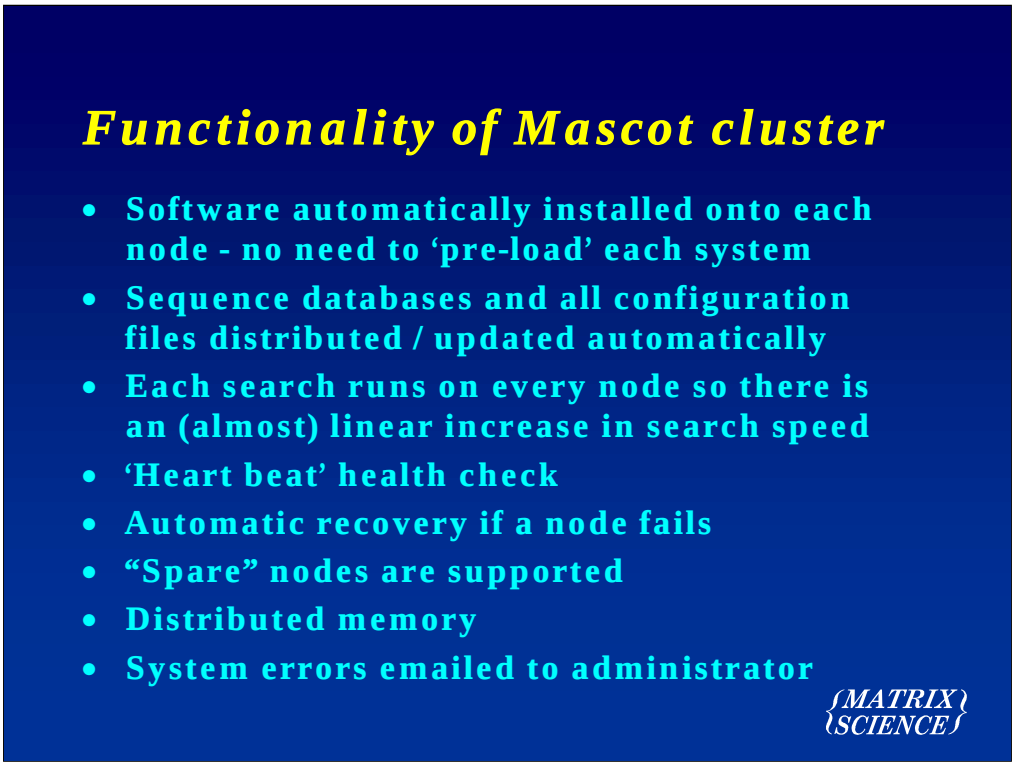
Setup under Windows is extremely simple



If the licence is for 4 CPU's or more, you will be presented with additional dialogs for configuring the cluster nodes

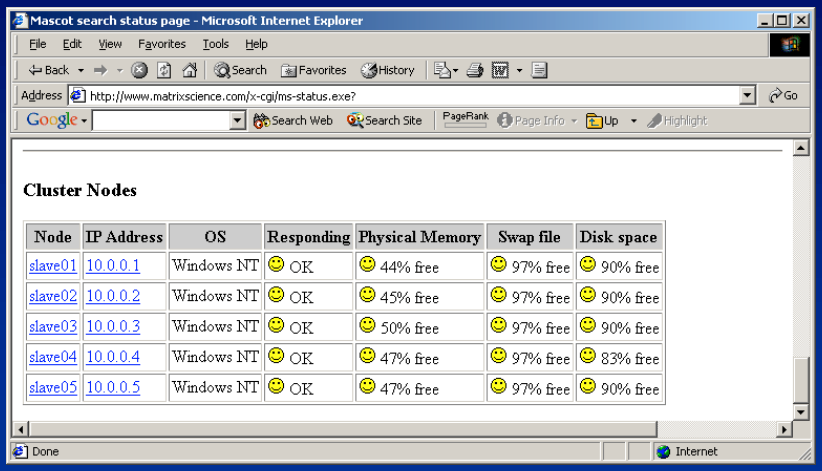


Under Unix, its just a case of editing two configuration files



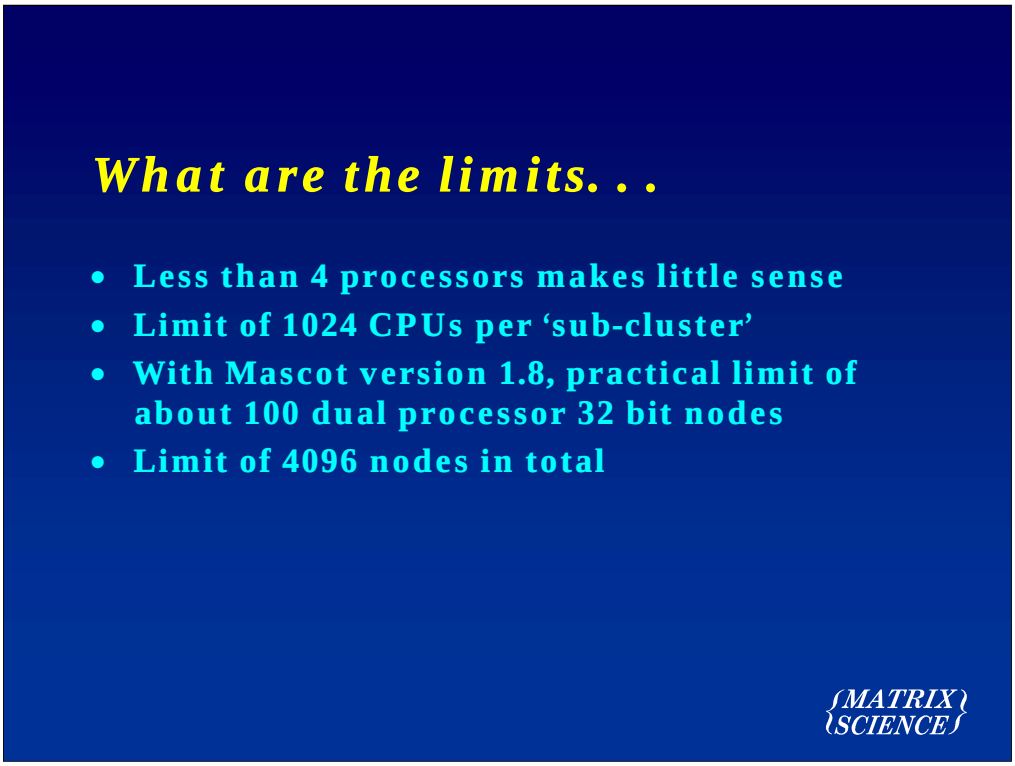
Cluster support includes many features to support large, high throughput systems

Status report

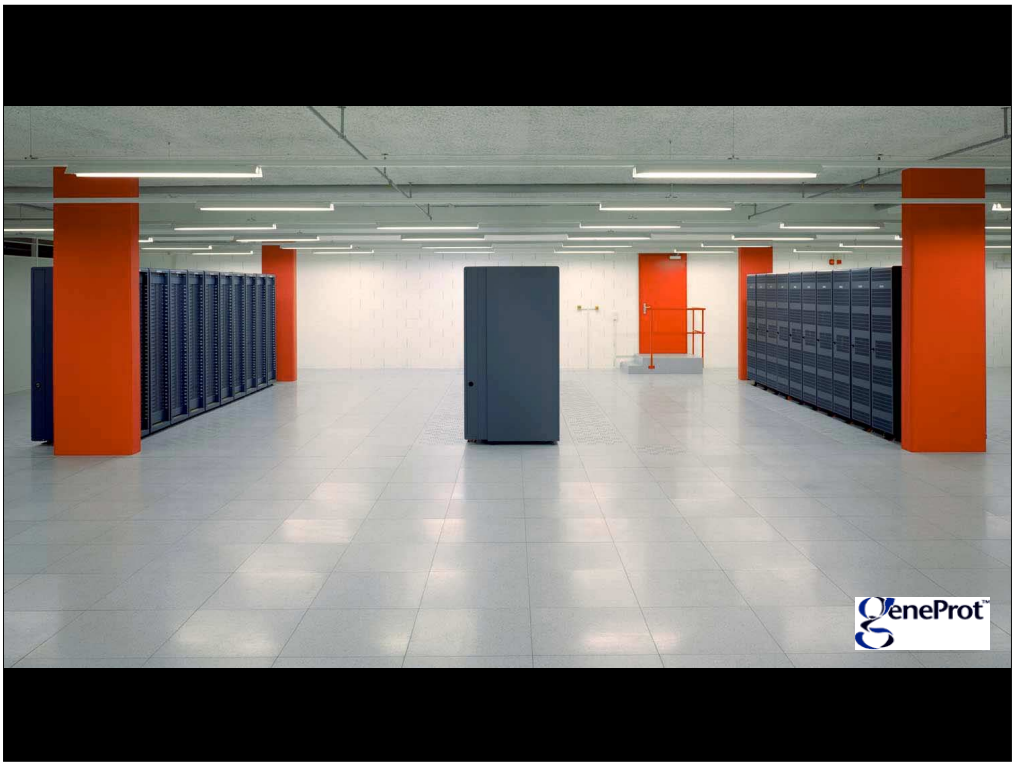


{MATRIX SCIENCE}

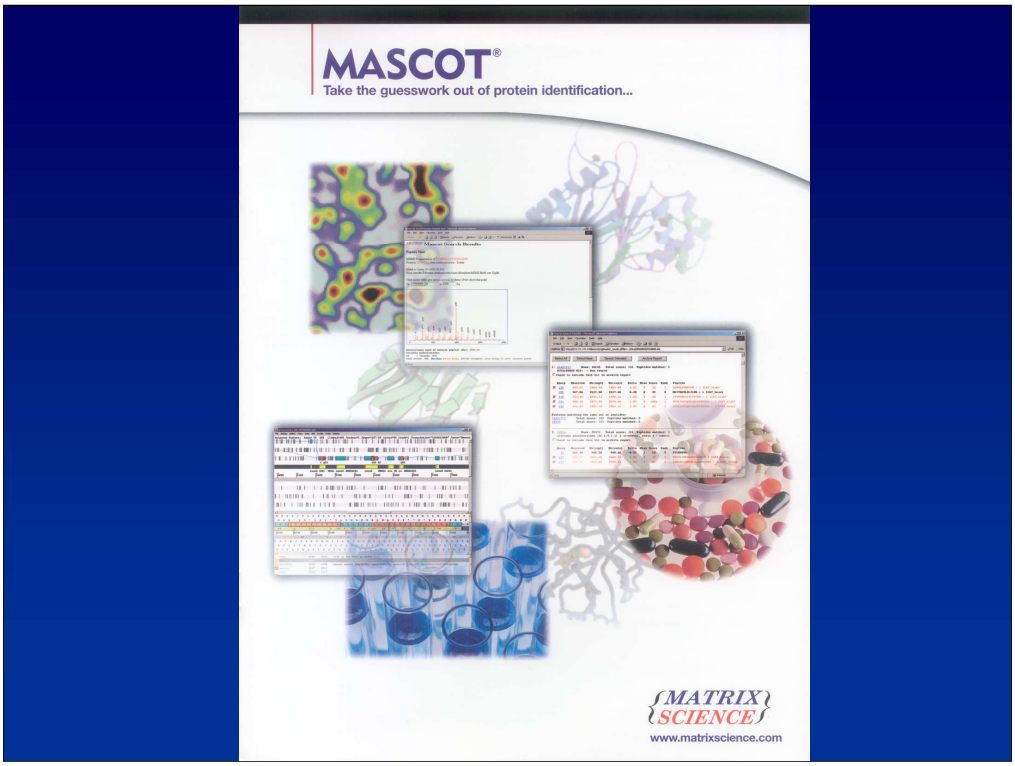
"At a glance" status screen



Of course, there have to be some limits



GeneProt have the largest Mascot cluster. 1440 Alpha CPU's configured as a number of sub-clusters


MASCOT®
Take the guesswork out of protein identification...
MATRIX SCIENCE
www.matrixscience.com