

Technology and cost trends

CERN T0 and CAF computing facility

Bernd Panzer-Steindel, CERN/IT

panzer@mail.cern.ch

V1.3 12. December 2007

The documents describes the price and technology trends of commodity computing elements with a special focus on the equipment installed in the CERN computing center over the last ~3 years. Based on these figures and the analysis of trends in computing technology I will also try to estimate the cost development during the next few years.

CPU resources

Technology

The major change during the last 2 years was the introduction of the dual-core technology (and now quad-core and so on...). The race for higher and higher processor frequencies stopped when the power and heat problems became essentially unsolvable. The solution was to spread the computing over more processors, each one running at a lower frequency and also reducing at the same time the leakage-currents with new techniques and materials.

http://www.xbitlabs.com/articles/cpu/display/core2extreme-qx9650_3.html

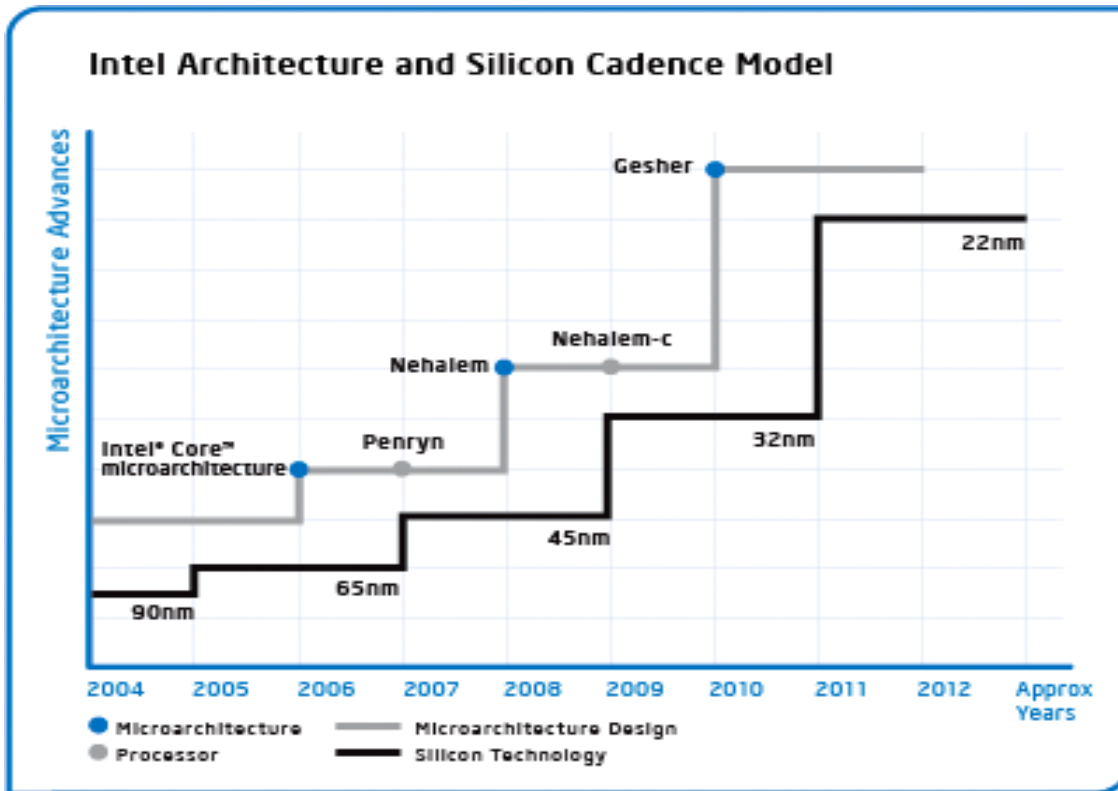
AMD was the first to get into the market with dual-core processors in 90nm technology during 2005. Intel followed during the middle of 2006 with their Woodcrest line in 65nm technology and has taken the lead since then.

At the end of 2005 Intel introduced the processor production on the 65nm scale and is just releasing the first CPU's (November 2007) with 45nm structure width. I will concentrate here on the Intel road maps, as they have a lead of more than 12 month in this technology compared to AMD who have just increased their production rate in the 65nm processors, e.g. their new flagship the Barcelona chip is just being announced and bulk shipping is expected by the first month of 2008.

The general Intel road map is now following their so called 'tick-tock' model :

<http://www.intel.com/technology/magazine/computing/cadence-1006.htm> ,

<http://www.edn.com/blog/400000040/post/1870014787.html>



The width of the chip structure is decreased every 2 years (tick) and there is an architecture change/improved interleaved also every 2 years (tock).

First prototypes of electronic components with 32nm wide structures have already been produced. A good overview of the future road maps and developments can be found here : http://www.tomshardware.com/2006/09/27/idf_fall_2006/index.html
<http://www.tgdaily.com/content/view/33998/123/>

The rate with which the transistors are shrinking seems to be ensured for the next ~6 years, but there are severe problems to overcome beyond the 22nm structures.
<http://www.edn.com/index.asp?layout=article&articleid=CA6461376>

There has been also a major change in the form factors for the complete PC box. Two half size motherboards are packed into a 1U chassis with a very efficient common power supply. From our measurement this solution seems to be even more power efficient than standard blade solutions. This type of packaging seems to become a strong trend and a strong competition to blade systems.



New processor generations are now deployed with larger level 2 caches which can improve the performance considerably. It is not clear how the HEP community can profit from this directly, as the majority of the reconstruction and analysis programs are still undergoing quite some changes and thus there are difficulties to provide exact benchmarking. A general investigation about cache sizes and their effects can be found here:

http://www.tomshardware.com/2007/10/24/does_cache_size_matter/index.html

Risks

The developments mentioned in the following points might have large effects in the general PC market and thus might affect the costs we (HEP) have to pay for the PC's needed to run our applications. The risks are always coming more from the market developments and not so much from the technologies per se.

Multi-core

There has been the paradigm change during the last 18 month to move from higher frequency single processors to multiple lower frequency processors. The chip production lines are setup to produce CPU's with many cores on the chip. Prototypes of 80 cores on a chip have already been presented by Intel : http://www.news.com/2100-1006_3-6119618.html

But it is not clear what the market directions really are and the paradigm change in hardware requires also a paradigm change in software. There are currently several independent developments ongoing in industry to exploit the different possibilities, which also show the uncertainties in this area:

1. The roadmaps for the future CPU's and the number of cores available is updated frequently. Does an 8-way core have a market change in the desktop commodity market ? There are now also plans to ship 'odd-numbered' CPU's with 3 or 6 cores per chip.
2. There is a heavy 'promotion' of multi-threaded programming ongoing. But this is a very complicated area and is not well established in the general software industry. The effort is put into providing improved C++ building blocks, having a better multi threading support in the compilers and even to provide a 'middleware' system to automatically parallelize code.
<http://softwarecommunity.intel.com/articles/eng/1276.htm>
3. A special case is the trend to combine much closer the CPU and the graphics processor (GPU). On the one side we have products like the FireStream

‘coprocessor’ from AMD, a product developed during the last year after AMD bought the graphics specialist ATI.

And we have the market leader in graphics processors, NVIDIA, who launched lately their Tesla product, a standalone ‘supercomputer’ based on several linked graphics cards.

Intel on the other side is trying to establish a model, where one of the cores is doing the graphics work. They have bought in September the company HAVOC which produces the widely used physics engine (software product) for PC games which would run specially optimized on another core on a multi-core chip. Intel is also pushing a ‘new’ paradigm in the graphics processing model. Today the rendering principles in games are based on so called shader units, which require the special graphic cards from ATI and NVIDIA. Ray-tracing is a much better approach to produce realistic environment, but this requires much more computing capacity, which multi-core chip could offer in the future.

Using graphics processors for scientific computing purposes is not a new trend (<http://www.gpgpu.org/>), the limiting factors were the single precision floating point units, the complicated programming model and the limited amount of memory. The latter is by the way the major point why the HEP community can’t use the very cost effective game machines, while they are used as cost effective supercomputer in some sciences (PS3 has 512 MB memory, <http://www.physorg.com/news92674403.html>)

4. The multi-core environment makes it easier to provide systems-on-a-chip. In this model one of the cores provides the role of the main processor while the rest acts as special engines for graphics, audio, video, speech, etc. processing, integrated co-processors. Research is going on to evaluate the possibility of using the cores as re-programmable hardware to adapt on –the-fly to the changing software functionality (keyword FPGA).

All these developments might lead to special processor setups focused on the consumer market, but not very well suited for HEP applications.

Notebooks

There is a clear trend away from the desktop PC moving towards the notebook.. The latest analysis of the markets show much larger growth rates in the notebook area.

http://crave.cnet.com/8301-1_105-9799162-1.html

<http://computers.tekrati.com/research/9536/>

As a consequence the focus will be more and more on mobile processors with their specific usage patterns in notebooks. Energy saving measures in this area are in general not beneficial for the server processors and the usage pattern in the HEP community. Today already 42 % of the Intel income comes from mobile sector.

AMD Intel competition

While in 2005 and 2006 the processors from AMD had clear advantages over the corresponding processor generations from INTEL, the picture has changed completely in 2007. After internal restructuring and a change in their strategies, Intel was able to push their new multi-core processor generations with much success into the market.

AMD was able to keep their market shares in the desktop, server and mobile area only because of a very aggressive pricing strategy.

Market shares → Overall 23.5 % for AMD and 76.3% for Intel (AMD mobile 18.9 %, AMD server 13.9 %)

The stock market shares of AMD have dropped by a factor two during the last 12 month.

This strategy caused heavy losses for AMD during the last 4 quarters (more than 2 billion \$) and AMD is from the technology point of view still about 12 month behind Intel.

In addition they are late in introducing their newest processor lines (Barcelona, Phenom).

The risk is that AMD has to scale down and becomes less of a competitor for Intel with the corresponding consequences for the pricing strategies (slowdown in cost decrease).

Quarterly results for revenues and profit in billion dollars:

	Q1/05	Q2/05	Q3/05	Q4/05	Q1/06	Q2/06	Q3/06	Q4/06	Q1/07	Q2/07	Q3/07
INTEL Revenue	9.4	9.2	10.0	10.2	8.9	8.0	8.7	9.7	8.9	8.7	10.1
INTEL Profit	2.2	2.0	2.0	2.5	1.4	0.9	1.3	1.5	1.6	1.3	1.9
AMD Revenue	1.2	1.3	1.5	1.8	1.3	1.2	1.3	1.8	1.2	1.4	1.6
AMD Profit	-0.02	0.01	0.08	0.96	0.19	0.09	0.14	-0.57	-0.61	-0.60	-0.40

Table 1 Revenues and profits for AMD and Intel over the last 2 years

Costs

The following plot shows the evolution of the processor costs during the last 3 years.

The reference value for the performance calculations is the 2.8 GHz Intel Xeon processor with 2000 SI2000. There are several interesting things to be noted in the plot.

- There is about a factor 2 price difference between the ‘single’ processor version and the DP version (capable of running in a dual socket configuration). The actual difference in the architecture and chip layout is very low and would not justify this difference, thus it is a pure pricing policy.
- The quad core systems got a large boost in terms of price/performance during the second quarter of 2007. They became sometimes even cheaper than the dual core systems. Again a pure pricing policy to get even further ahead of AMD.

The green curve shows the calculated price/performance of a complete server box (chassis, power supply, motherboard, 2 GB memory per core, one disk per 4 cores, dual-CPU 4 cores each). This is based on the outcome of the regularly issued CERN tenders. The extra costs for the rack infrastructure, networking and console infrastructure are not included, which in general would be about 5-10% of the box costs. But this is of course very specific to the computer center layout.

Processor cost evolution

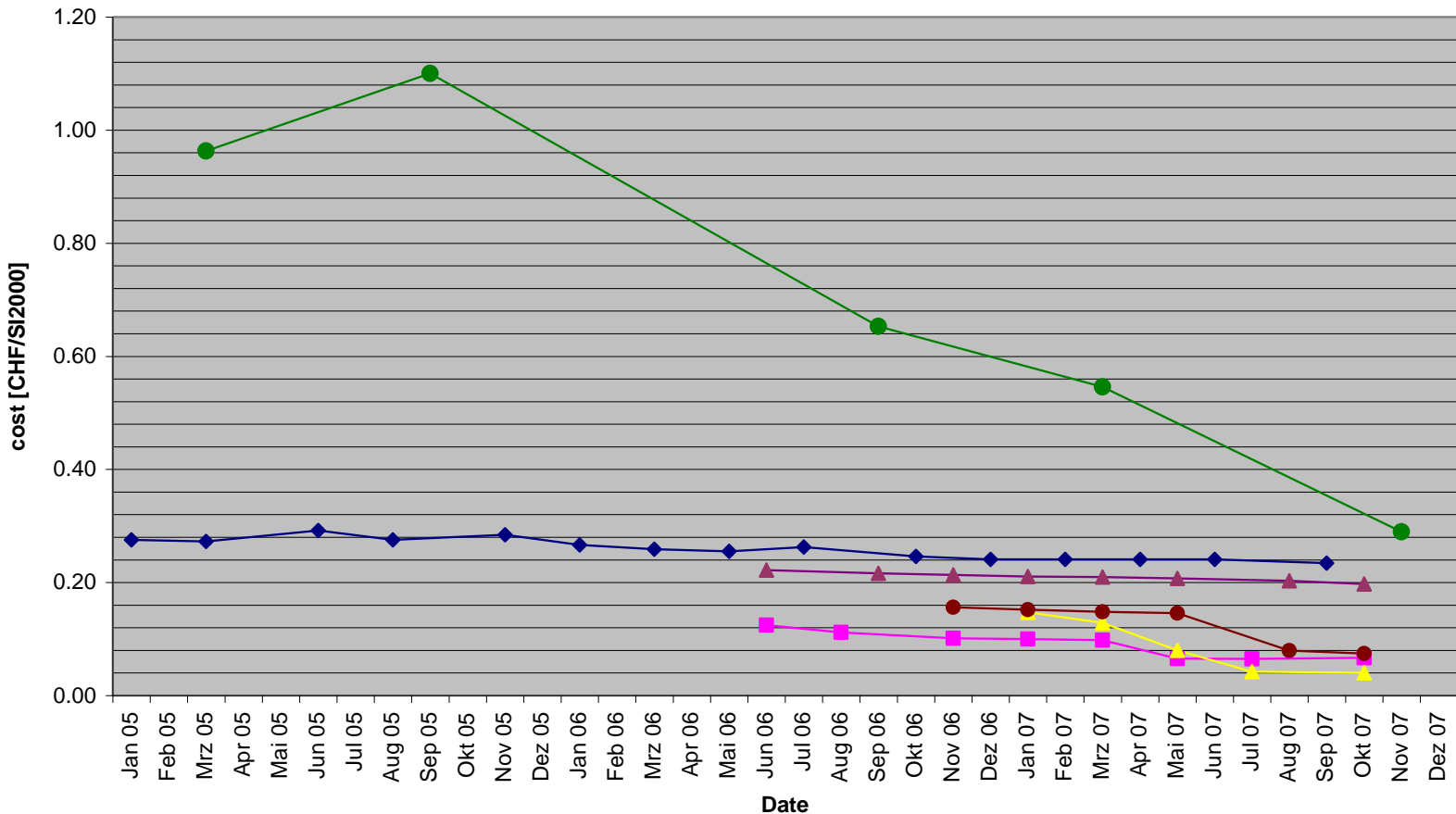
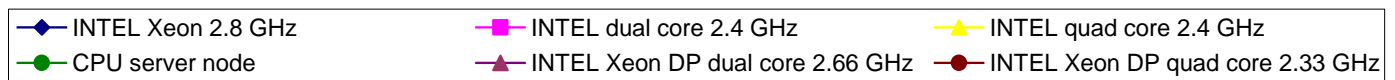


Figure 1 *Cost evolution of processors and PC's over the last 2 years*

The component prices are taken from a European wide price comparison web side (<http://www.heise.de/preisvergleich/eu/>) and than transformed from Euro into Swiss francs based on the exchange rates during the last 3 years (average was about 1.65 CHF/Euro). In addition the prices are reduced by 20% because CERN in not paying any VAT costs.

The following table shows the predictions of the cost evolution which was made in 2005 :

Year	2004	2005	2006	2007	2008	2009	2010
CHF/SI2000		1.57	1.05	0.7	0.46	0.31	0.21

Table 2 *Old cost predictions for CPU capacity from September 2005*

These numbers have to be corrected taking into account the real costs CERN paid for the 2008 deliveries.

year	2008	2009	2010	2011	2012
Cost [CHF/SI2000]	0.30	0.22	0.16	0.11	0.08

Table 3 *New cost predictions for CPU capacity from October 2007*

While the difference between the predicted 2007 costs and the real 2007 costs was only about 15 % (better than expected), there is a major improvement for the 2008 values.

This is due to the introduction of new multi-core systems by Intel and a heavy price war between Intel and AMD.

Taking the mentioned risks into account we assume now that the costs are improving by a factor 1.4 per year.

Memory

Today the majority of memory modules are based on DDR2. The next generation with improved memory speed has already been introduced into the market (DDR3), but still with a very low share (1-2%). This will increase over the next 2 years. A concise analysis of the technology and the markets can be found here :

http://www.tomshardware.com/2007/10/03/pc_memory/index.html

In the high end graphics market companies have already started to introduce DDR5 memory.

The new memory type offers improved speed , but at the cost of lower latency. The effect of memory speed and latency on the HEP code is not very clear. A recent analysis of the CMS code shows that a lot of time is spend on load+store operations.

<http://indico.cern.ch/getFile.py/access?contribId=8&sessionId=0&resId=0&materialId=slides&confId=15918>

Since several years the industry is working on non-volatile memory technologies (MRAM, FeRAM, PRAM). Despite the different announcements and quite some progress in the research area (Toshiba, IBM), only little and expensive memory module of this type have been introduced into the market (Freescale with 4-Mbit MRAM modules).

The HEP community requires about 2 GB of memory per core to be able to run the different reconstruction, calibration and analysis jobs efficiently on the PC's. Thus a dual-cpu quad-core node needs 16 Gbytes of memory. Thus memory contributes about 30% to the overall PC costs and about 35% to the overall PC power consumption.

The cost evolution of memory shows in general a very good decreasing trend, but still not in a smooth manner, as can be seen in the following figure:

Memory cost evolution

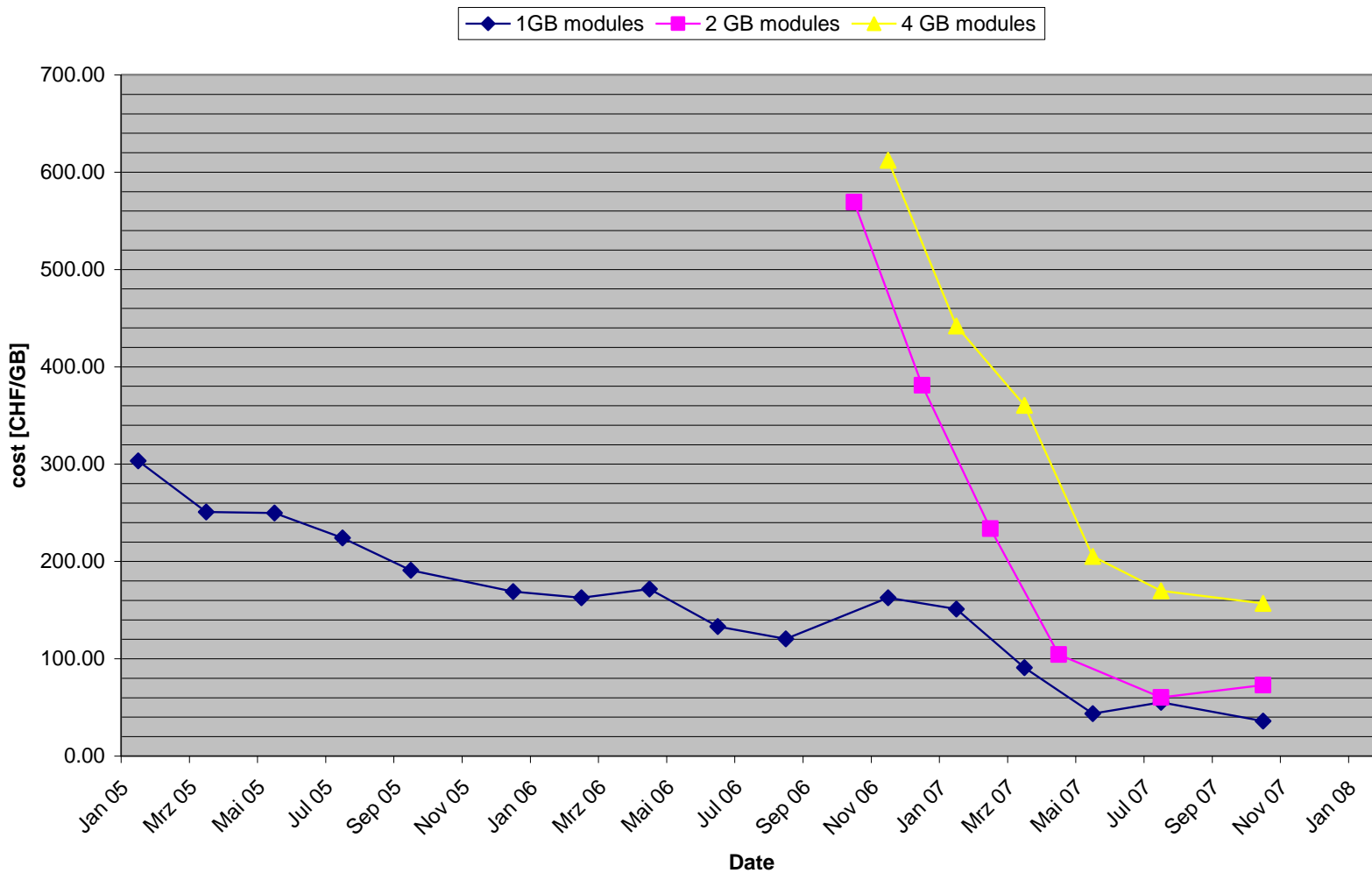


Figure 2 Cost evolution of DDR2 memory modules (667 MHz) over the last 2 years

Disk resources

The market for hard disk is still growing constantly and expected to continue to do so over the next years. The total shipment of disks in 2007 is expected to reach 500 million units (~ 4500 PetaBytes of space) and create 30-35 billion \$ revenues

<http://www.integratedmar.com/ereportsb/story.cfm?item=462>

The space density growth rate per year is on average about 40%. The latest recording technology (PMR, Perpendicular Magnetic Recording) has reached already a 60% market share in 2007 and is expected to dominate the markets at 100% in 2008. There are still major technology breakthroughs on a regular basis, the latest one enables a quadrupling of the current density which will lead to 3-4 TB disks by 2011.

<http://www.physorg.com/news111637180.html>

The domination form factor in the market are still 3.5 inch disks, but the large growth rates in the mobile market have lead already to a share of 35% for disks with a form

factor of 1.8 and 2.5 inch. This is expected to grow to 44% in 2012. The maximum amount of space available on a 2.5 inch disk is today 250 GB, while it is 1 TB for a 3.5 inch disk. The 2.5 inch disks are about a factor 2 more expensive.

There is a growing market for Solid State Disks, especially aimed at the mobile area. A very good overview of the market state of SSD disks can be found here:

<http://www.storagesearch.com/ssd-buyers-guide.html>

SSD disks are currently about a factor 60 more expensive than magnetic hard disks (CHF/GByte). But this factor was more than 200 about 15 months ago, so there has been a very steep cost reduction in a short time period.

The Predictions claim that SSD disks will in 30-40% of all notebooks by 2012.

More and more hard disk vendors are offering now so called hybrid disks, which have large flash memory in front of the physical disk layer (hundreds of MB up to multi GBytes). This comes along with a price offset of currently more than 25% compared to a 'normal' disk. The HEP applications and access patterns in general cannot profit from this cache and if this becomes a large scale, additional costs have to be taken into account.

During this year holographic storage made a comeback in the form of a product announcement from InPhase. They will sell moderately sized holographic disks with moderate performance but with a high price tag.

<http://www.computerweekly.com/Articles/2007/09/21/226902/holographic-data-storage-the-next-big-thing.htm>

Costs

The largest uncertainty is coming from the coupling of the disk storage to the application layer and the tape layer. The experiment have given their requirements based on the amount of RAW data and derived data (ESD, AOD). This does not take into account any access patterns and performance requirements.

- how many streams per spindle are reasonable ? ratio of CPU cores per spindle in the disk storage layer.
- Is RAID5 the right choice (rebuild time, data integrity checks) ? or RAID1 ?
- replication of data to remove hot spots are not included in the storage calculations
- the optimal usage of tapes requires careful load balancing and dedication of spindles for either read or write operations. To ensure an overall throughput more spindles are needed.
- The average access time on a disk has not improved since more than 5 years (8-10 ms for commodity disks), while the space per disk has increased considerably. As access scales with the number of spindles one has to buy automatically more space than 'needed'.
- Small files are causing many problems in the tape layer, thus the experiments have started to deploy large scale merging procedures. This of course needs more disk space for an additional cache layer and because of the many streams the cache has to handle.

The following plots shows the price performance development of single hard disks during the last 3 years. In addition the costs of a full server is also added. This typical disk server configuration comprises a multi-core CPU node with 8 GB of memory and 20 data disks configures as 3 RAID5 file systems, each with 1 'parity' disk and one spare disk. The space is always quoted as usable space, after subtracting the RAID5 and file system overhead. Again costs for infrastructure are not included (racks, power distribution, network), which depend heavily on the center setup (allowed power and cooling density and network blocking factor).

Disk cost evolution

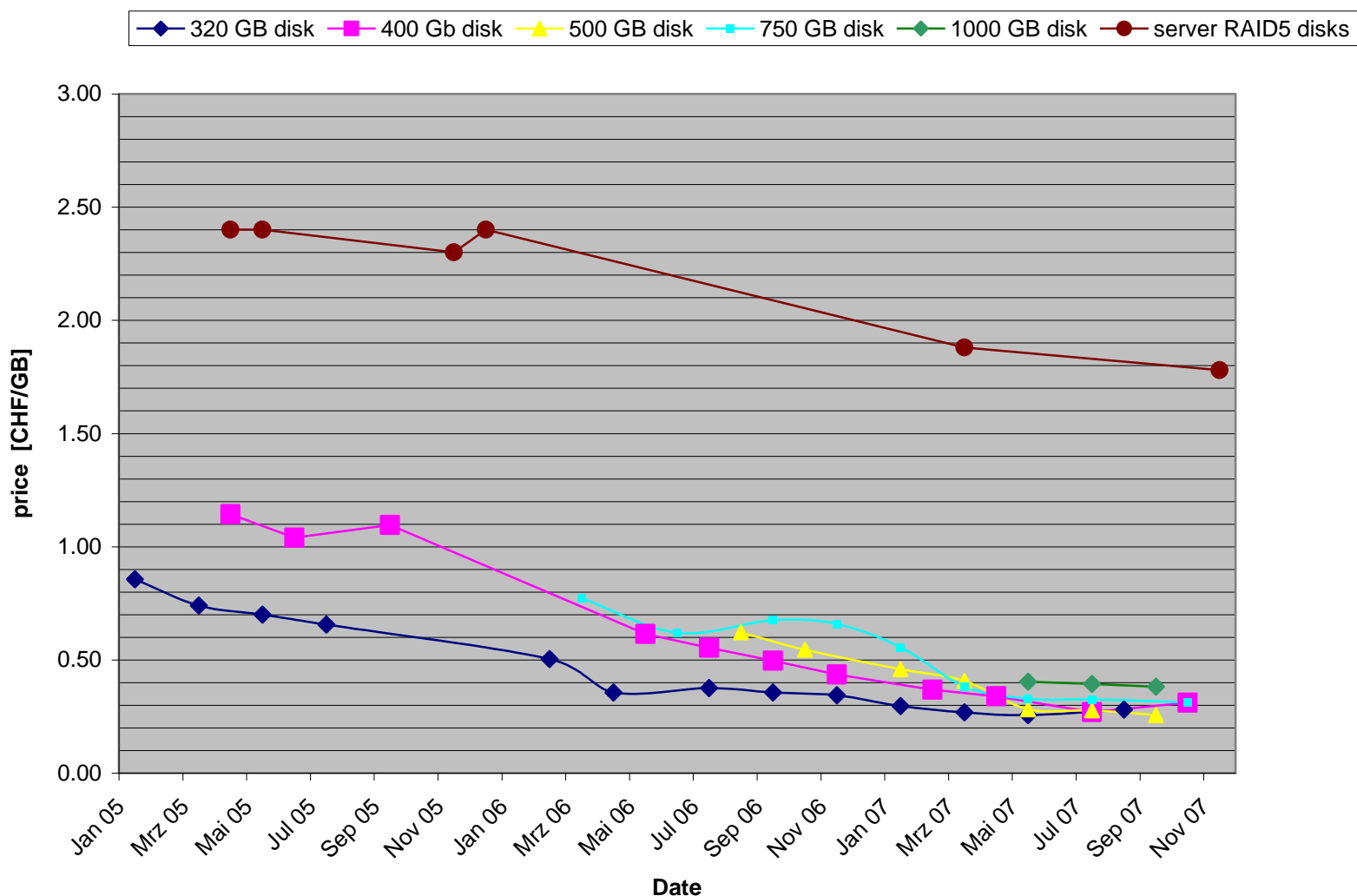


Figure 3 Cost evolution of single disks (SATA) and full disk servers

The component prices are taken from a European wide price comparison web side (<http://www.heise.de/preisvergleich/eu/>) and than transformed from Euro into Swiss francs based on the exchange rates during the last 3 years (average was about 1.65 CHF/Euro). In addition the prices are reduced by 20% because CERN in not paying any VAT costs.

The following table shows the predictions of the cost evolution which was made in 2005 :

Year	2004	2005	2006	2007	2008	2009	2010
CHF/GByte		4.5	3.75	2.93	1.83	1.14	0.72

Table 4 *Old cost predictions for disk capacity from September 2005*

There has been a large g improvement between 2006 and 2007 and the measured price performance value was 1.9 CHF/GB for 2007. This improvement did not continue for the 2008 deliveries, because of several boundary conditions we imposed on these disk We requested not the highest possible capacity per disk (max 500 GB) to maximize rthe number of spindles and increased the number of processors and memory per box to cope with higher application footprints and to be able to run checksums on a large scale in the future.

For the future predictions we assume an improvement factor of 1.3 per year.

Disk	2008	2009	2010	2011	2012
Cost [CHF/GB]	1.6	1.23	0.95	0.73	0.56

Table 4 *new cost predictions for disk capacity from October 2007*

Tapes

Despite reoccurring predictions about the ‘end’ of tape storage http://www.byteandswitch.com/document.asp?doc_id=135764 this market is still growing. The major player is the LTO consortium which just recently celebrated quite remarkable milestones. During the last 7 years about 2 million drives were sold. The important point here is that they sold during the period of Sept. 2006 until Sept 2007 about 500000 drives after the introduction of the LTO4 version (800 GB cartridge, 120 MB/s max speed). In addition about 80 million cartridges were sold in the last 7 years, but here again major part of this happened during the last 12 month (30 million cartridges, which corresponds to about 12000 PB of space compared with 4500 PB of disk space sold during the same period) http://www.byteandswitch.com/document.asp?doc_id=133041

The LTO technology had a market share of about 70% in 2006.

Also the cost per GB has dropped in a linear manner during the last years. The current commodity market costs are about 0.13 CHF/GB for an LTO3 cartridge (400 GB) and 0.17 CHF/GB for the newer LTO4 cartridge (800 GB).

The list prices for cartridges from the high-end tape drive sellers like IBM (3595 , 700 GB cartridges) and STK (T10K, 500 GB cartridges) are about 0.3 CHF/GB. Both vendors will introduce 1 TB tapes and the corresponding drives sometimes during the second part of 2008, which will cause a drop in the relative tape costs
A large robot for ~10000 tapes would cost about 0.5 MCHF.

Extrapolations and detailed costs analysis in this area is very difficult as large tape storage installations require large robotic installations from single vendors. The pricing policy of tape drives, media and robotics are than strongly coupled and purchased in a bundle. The total costs are strongly negation dependent.

Can tape be replaced with hard disks ?

The following back-of-an-envelope calculation tries to give a status of this question.

A tape storage system of 10 PB of data with an optimistic data transfer speed of 1 GBytes/s would cost today about 2.5 MCHF using LTO4 .

- 625 KCHF for 12500 0.8 TB cartridge slots
- 1500 KCHF for 10 PB of cartridge space at 0.15 CHF/GB
- 400 KCHF for 20 tape drives and the corresponding servers
(assuming 50 MB/s per drive)

A disk system should have similar characteristics and one cannot use the standard disk servers as a comparison. One can use a model where a simple CPU server has 100 disks connected via USB2. This has been proven to work on a small scale (20 disks) already in a small R&D investigation in 2006 at CERN.

- 800 KCHF for 200 CPU server including network and USB2 infrastructure
- 2800 KCHF for 23000 0.5 TB USB disks, all disks would be in a RAID5 configuration plus a few sparer disks per node

The total cost of such a disk system would be about 3.6 MCHF and could deliver 20 GBytes/s performance.

Today a disk system would still be a factor 1.5 more expensive than a corresponding tape storage system. But if one takes into account the price developments of disk and tape over the years, this difference of 1.5 would drop to about 1.1 by the end of 2008.

Summary

The overall developments in the CPU, disk and tape area are positive, both from the technology point of view and from the cost point of view.