Markets, Technology and Efficiency

Some ‘pseudo-random’ walk

Bernd Panzer-Steindel, CERN/IT

Expected 2011 market figures

Mobile phones 1650 million units

— Smartphones 480 million units (+60% CAGR)

Tablets 50M (+200%)

Netbooks 20M (-20%)

Notebooks 180M

Desktops 150M

PC market
350 M units
200 B$ revenues
-3% in Q1 2011

Server 10M (+12%)

- HPC servers

Server market
55 B$ revenues (HPC = 10B$)

We are in the low-end server market (dual-processors, ECC memory, BMC control, 12 memory DIMM slots)

20% error on the numbers, Gartner iSuppli, IDC report different numbers

Trends I

Strong push of the ARM processor --> smartphones, tablets
and integration of ARM into large scale servers --> low-end servers
Windows 8 support for ARM

AMD is pushing the 'new' heterogeneous architecture (CPU-GPU) into the desktop market and the HPC market. it's essentially the return of the co-processors (graphics, video, audio, encryption - efficient, low-power, specialised hardware for specific tasks)

INTEL with their MIC design and smaller companies like Tilera,SeaMicro, Quanta are trying to establish many-core systems in the HPC market (and the low end server market). Very difficult programming model to achieve high efficiency.
Increased focus on vector processing in the current and next processor generations.

Google, Facebook and Microsoft are designing their own servers (motherboards, power supplies) for the efficient use in their mega-data centres

Large smartphone and tablet growth rates and moderate server growth rate
--> model: low power devices using cloud services

Trends II

Low growth rates int the desktop area and lower than expected growth rates int the server market are due to several factor:

1. **upgrades were already done on a larger scale in 2010, postponed from the recession year 2009**
2. **increased efforts for higher efficiencies by using virtualization on all levels**
3. **optimising TCO by combining low end devices with cloud services**

The large scale trend to smartphones, tablets and partly notebooks has created some side-effects:

1. **the magnetic hard disk growth rates have been cut by a factor 2**
2. **high demand for flash memory and SSDs. 22 B$/y market revenues price increases are expected soon**
3. **SDRAM prices are still falling, 33 B$/y market revenues**

Trends III, processors

low power processors with ~2 cores

many-cores (50-core INTEL MIC, 64-core Tilera) for HPC and specialised tasks

multi-core processors (6-core INTEL , 12-core AMD)

INTEL share of the market is 83% in the desktop and 92% in the server area

the current AMD 4-CPU 48-core systems are exceptions, subsidised for the HPC and server market, low performance cores, AMD will change architecture in 2012/2013 and follow INTEL strategy (tick-tock)

INTEL strategy for the multi-core system has changed. moving from a geometric to a arithmetic increase in cores. instead of doubling now +2 per year.

2011 2 * 6 cores (Westmere) plus 25% from SMT = 15 cores per node
2012 2 * 8 cores (Sandy bridge) plus 25% from SMT = 20 cores per node
2013 2 * 10 cores (Ivy bridge) plus 25% from SMT = 25 cores per node

(depends also on the alignment of technology releases and purchasing cycles)

Trends IV, hard disks

Major vendor consolidation, only three left, two dominating: Seagate and Western Digital

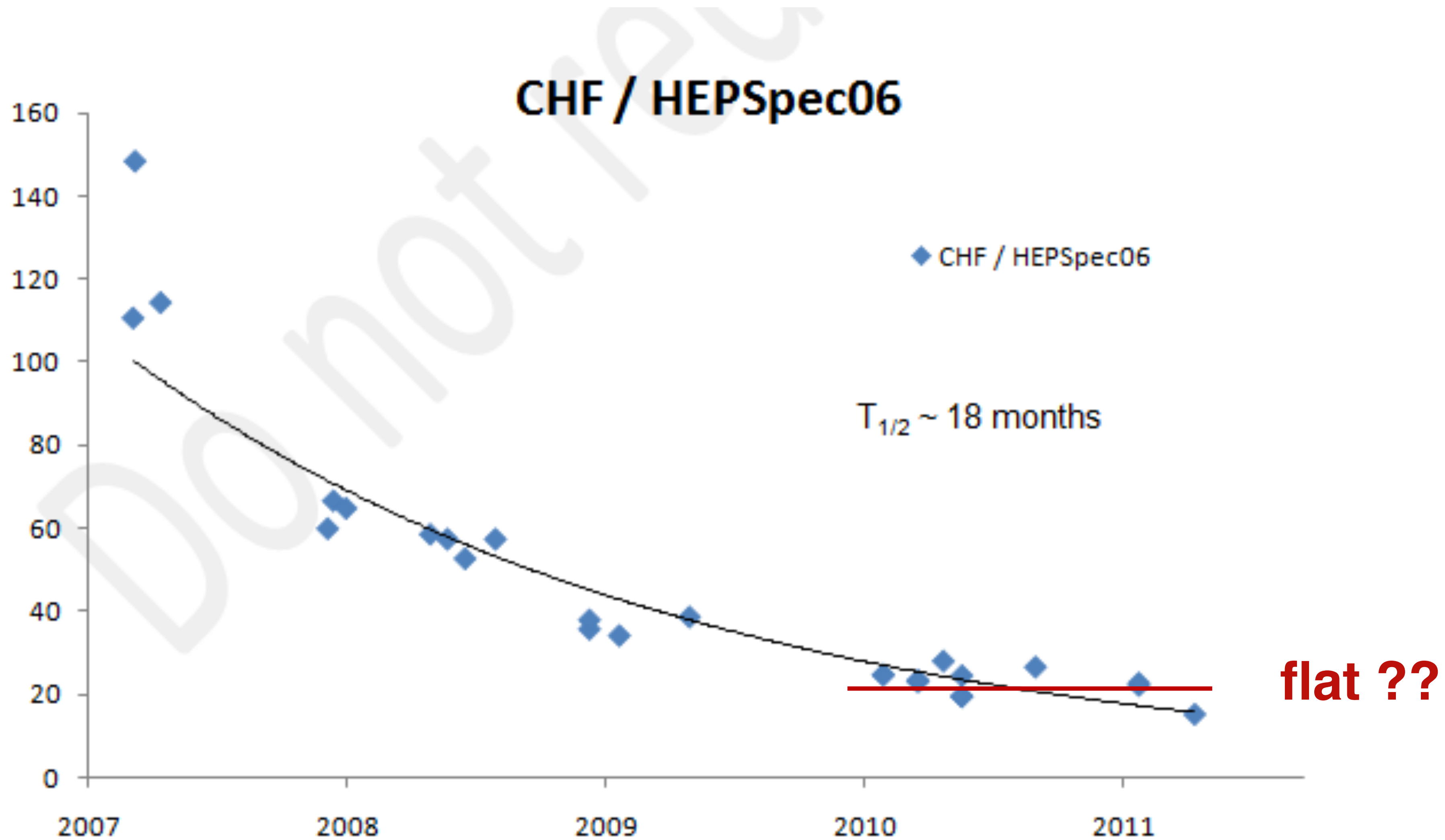
Increasing market pressure on magnetic hard disks from SSDs
35 B$ HDD market, 4 B$ SSD market

Competition and fast technology cycles are causing reliability issues.
we have now regular large scale replacement campaigns of disks.
--> software implications to maintain efficiency: fault tolerance of servers, replication of data (maybe need to go to 3 replicas instead of 2)

Next generation of hard disk technology HAMR (heat assisted magnetic recording) and/or BPM (Bit-patterned media) are very complicated and expensive. Strong collaboration of the vendors required and established.
--> density growth rate slowdown expected

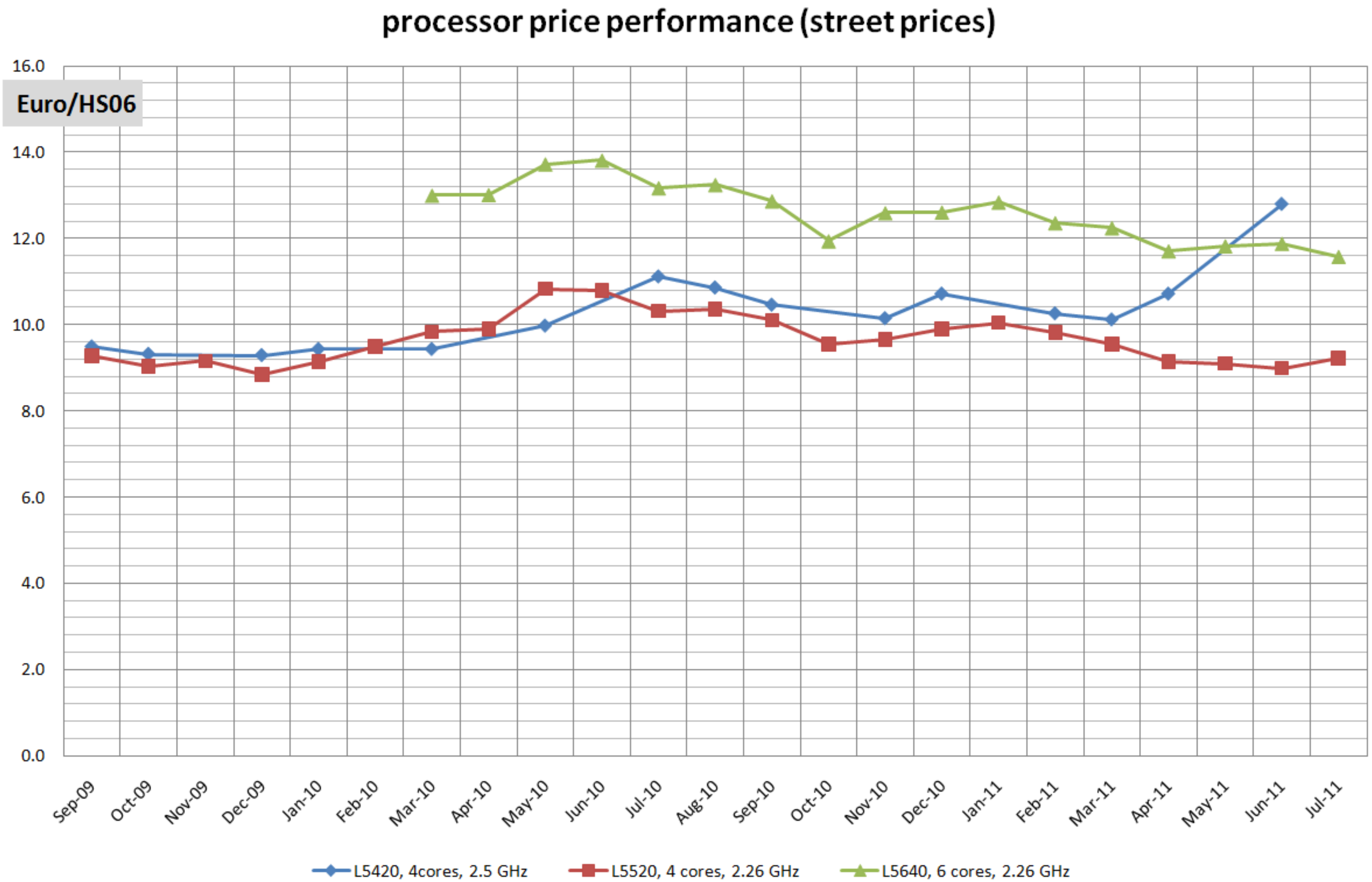
Processor price performance I

the observed price/performance curve at CERN seems to still follow an exponential decrease



the 2011 numbers define the cost for the 2012 equipment, cost shift by one year due to long purchasing cycle

Processor price performance II



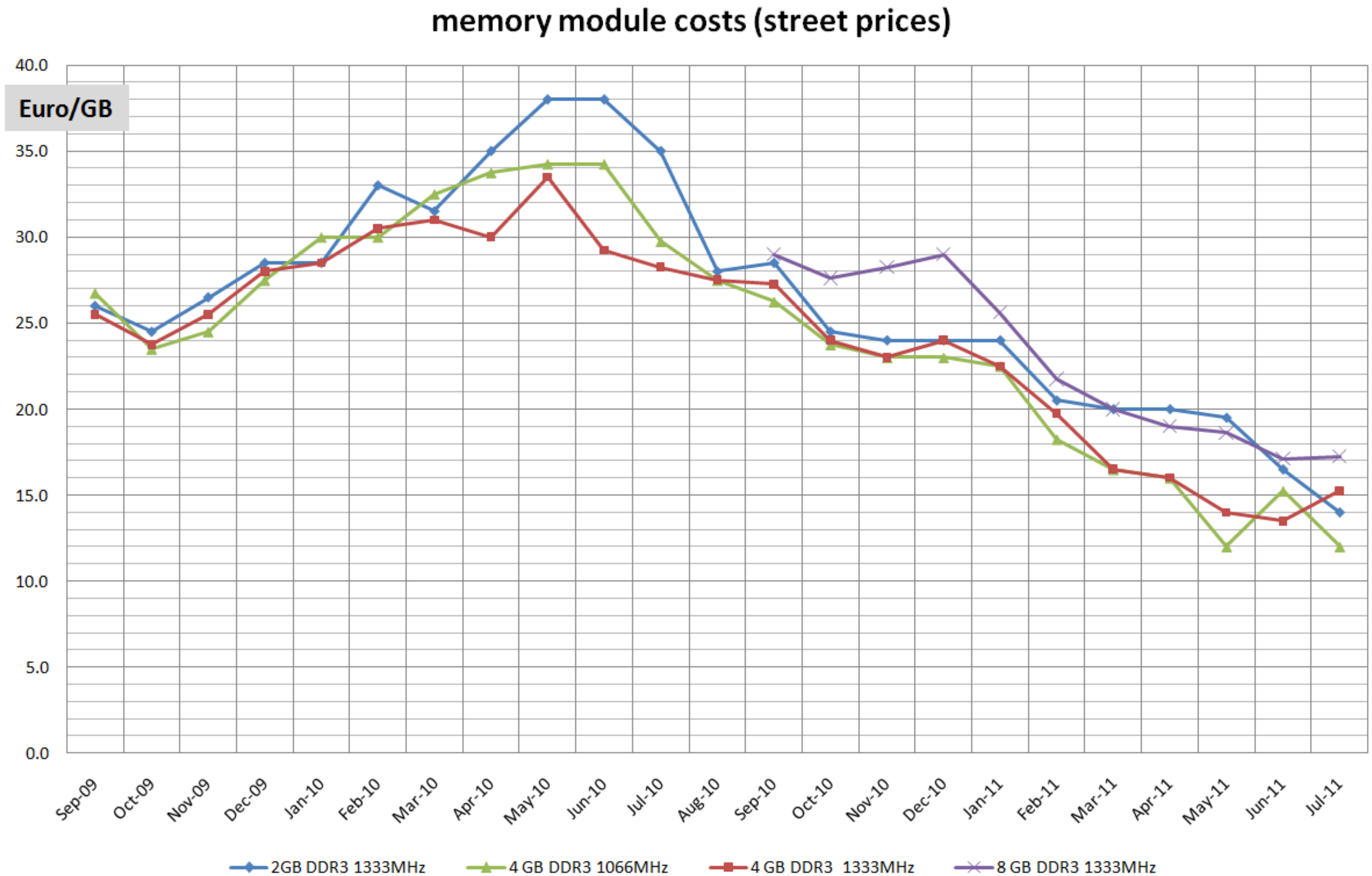
Moore's Law is about structure density and not about cost

the performance per processor is increasing, the performance per core stays constant or is even decreasing

price/performance over time is flat

the price/performance per processor is actually <u>increasing</u>

Memory price performance I



memory price/performance is 'bumpy' and the decrease more linear than exponential

but the decrease is essential for the node price/performance improvements --> ratio of processor share to node infrastructure

Efficiency

computing equipment growth reduction due to large efforts to increase efficiency in industry

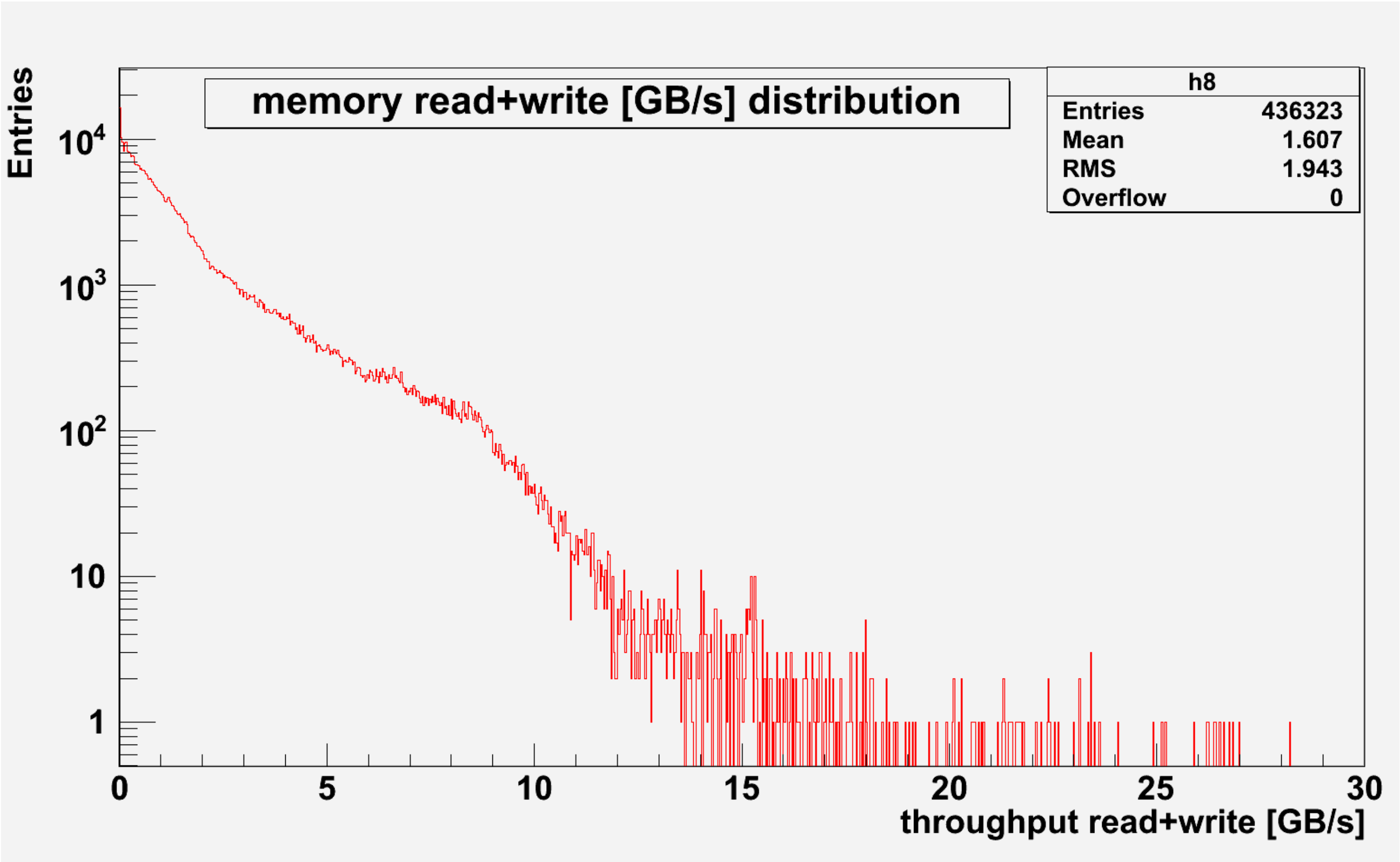
at CERN we have over-commissioned the installed hardware to cope with unforeseen problems and software deficiencies

computing works well, thus it is the right time to look in depth into improving efficiencies

requires investment into monitoring, debugging and understanding complicated cross-correlations:
batch-storage levels- software repositories-configuration management-network-experiment frameworks-.......

Efficiency, memory

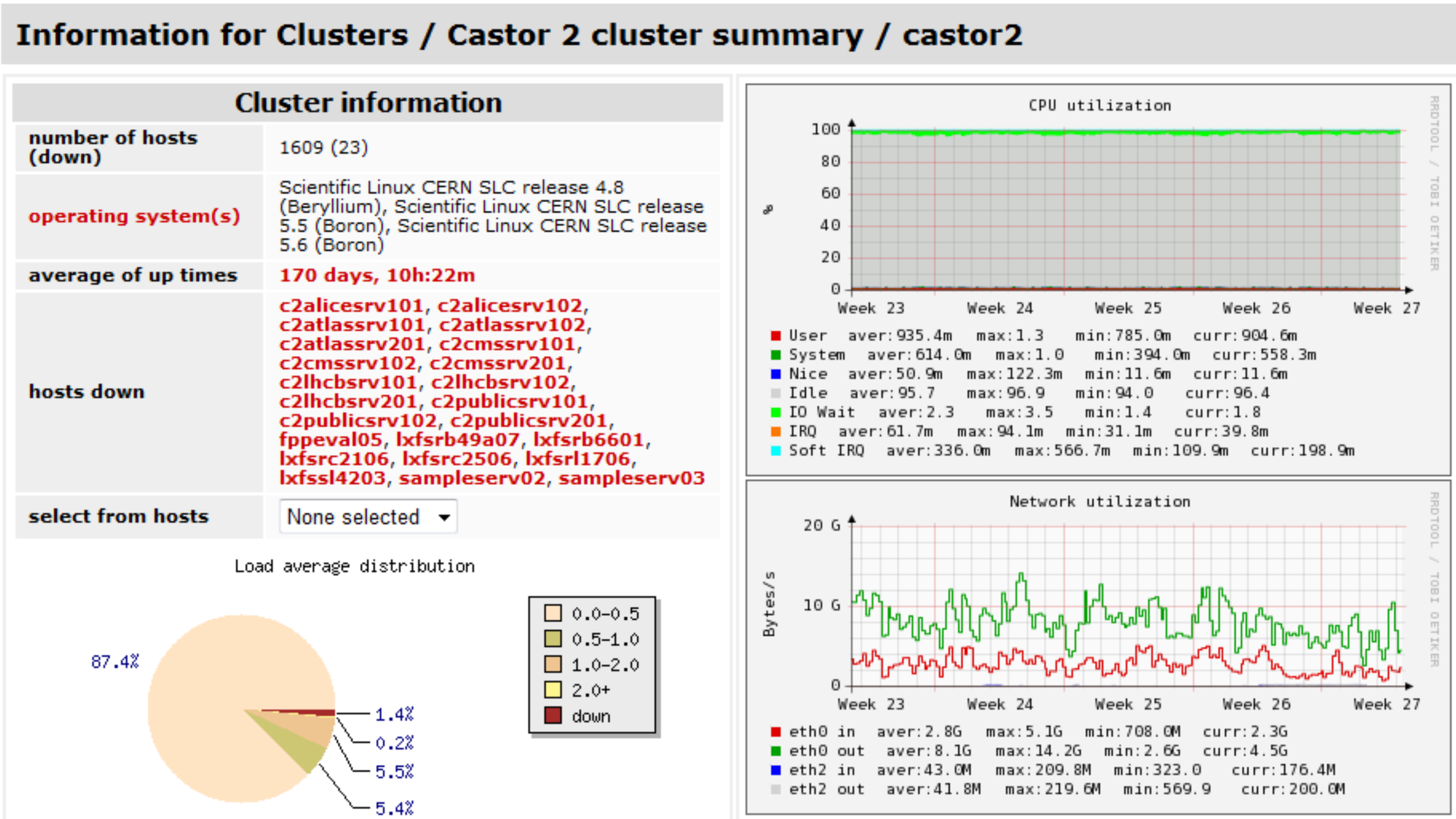
measured aggregate memory performance of jobs on an 8-core system



there is actually no problem with memory speed, >factor 10 safety margin on average

could optimise memory DIMM frequencies, memory channels efficiency improvements probably at the <10 % level

Disk server efficiency I



30 MB/s == 10% of one core

aggregate data rates of 20 GB/s
large amount of space compared to the needed IO performance, less than 5% CPU usage (disk server have single processors)

Disk server efficiency II

CPU usage of disk server is very low <5%

several possibilities to increase the overall efficiency of the disk storage system:

1. **Could add in principle ~15-20% CPU resources to batch by running jobs on disk servers**

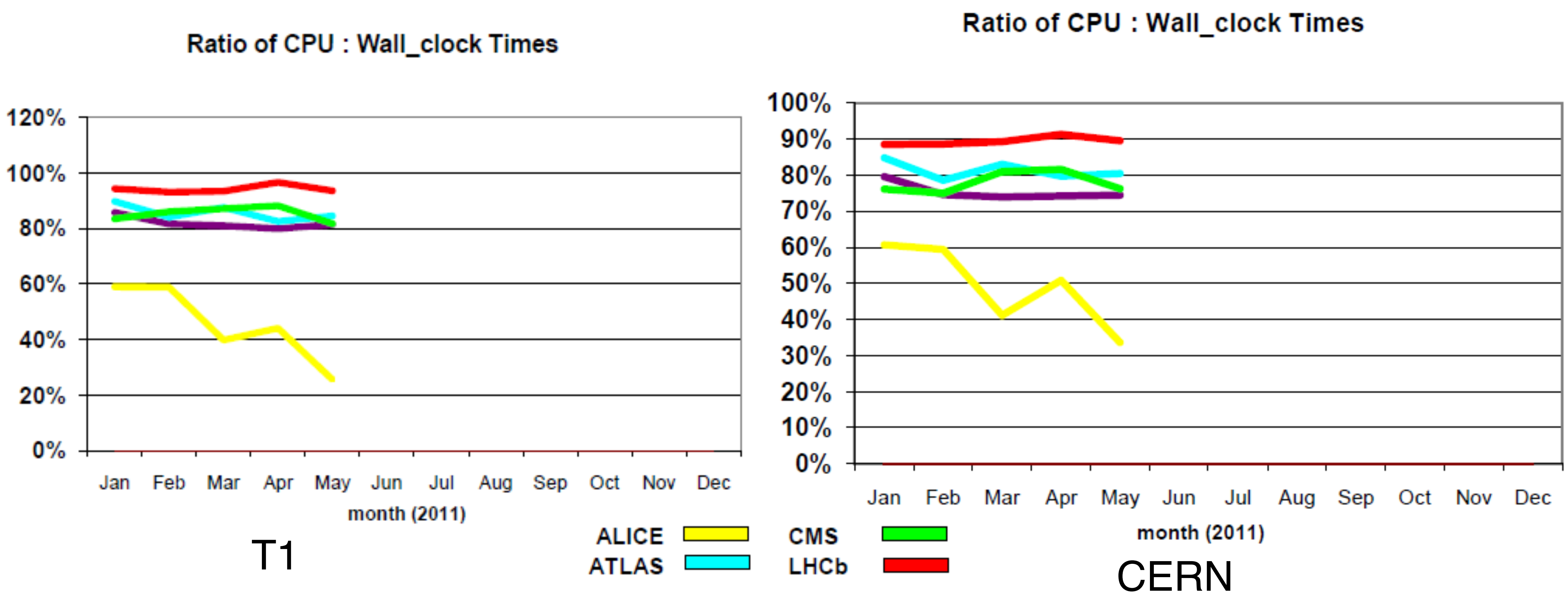
2. **merge CPU and disk servers, e.g. multi-core system with 6-10 disks, simple controller**

3. **larger disk servers, currently 24 disks per server --> 36 or 60 disks using low end processors (draining and filling takes long time, 10 Gbit required)**

Requires capable data management software

side-effects on whole node scheduling....

Job efficiency



for reconstruction and Monte Carlo the intrinsic system IO overhead (wait time) is of the order 1% per job

requires more monitoring and debugging, fine grain IO profiles of job categories, early warnings, etc.

CPU server efficiency I

Example list of possible sources for in-efficiencies:

- **Production state (new installation ,burn-in, failure rate) --> 80% varying**
- **Slot efficiency (dedication, user/queue limits) --> 5 - 95%**
- **Job efficiency --> 40 - 90% eff**
- **Node crashes and reboots (software updates) --> 99 % eff**
- **Stop of jobs due to wrong specs (queue limits) --> 98% eff**
- **Bad user jobs --> not monitored**
- **Processor technology matching of code, experiment code = 0.5 instructions per cycle today's processors can do 40 I/C, technology move to vector processing (SIMD) --> 3% eff**

problem: identification of the various efficiency effects and their measurement, various ways to improve --> side effects

Worker node storage

copy files to worker node or read directly from repository

spikes of high sequential IO plus background low IO
different usage , VO and job category dependent

more spindles needed as more jobs run
currently 2-3 x 3.5" --> more disks or SSD, space versus IO

cost factor, where to optimise ?
multi VO plus shared batch helps, mix job categories to spread workload, IO overhead low : 0.4 cores for 120 MB/s

--> coupled to CPU-disk optimisation, whole node scheduling
space for VO infrastructure, virtualization, disk-less nodes ?!
configuration management improvements
implementation of cloud/S3 storage

can we actually agree on a common strategy ?!

Whole node scheduling issues I

Whole node scheduling == dedicated resources
from experience at CERN: efficiency for dedicated Resources = 5-30%
efficiency for shared resources > 70%
(multi VO, job categories, mix and match, better 'Tetris')

possible need for extra system management --> core pinning of threads

IO access is similar to standard batch usage, but

- 12 jobs reading 12 different files on the storage system (= sequential) is not the same as 12 threads reading from the same file (= random)
- copy to worker node storage, possible merging of files

need to have solid proof of good efficiency before a major production deployment

Whole node scheduling issues II

Memory improvements

CMS : up to a factor 3 memory size improvements, event throughput the same as with non-threaded program

ATLAS : 25% memory improvements, ~ 10% loss of event throughput

--> creating heterogeneity between VOs

Cost estimates:

- reducing the memory by 50% would gain about 8% of the cost for CPU servers
- moving to 3 GB memory per core (physical and SMT) would add ~ 10% to the cost of a worker node

Summary

- mobile computing is driving the market
- nothing striking in the technology area, arithmetic increase of cores
- price/performance improvements could be slowing down
- started campaign to understand and improve overall efficiency at CERN
- common strategy for worker node file-copy !?
- whole node scheduling needs to proof good efficiency
- move to 3 GB of memory per core ?!