HEPIX 2006

CPU technology session

some ‘random walk’

INTEL and AMD roadmaps

- **INTEL has moved now to 65 nm fabrication**
- **new micro-architecture based on mobile processor development, Merom design (Israel)**
- **Woodcrest (Q3) claims + 80% performance compared with 2.8 GHz while 35% power decrease**
  some focus on SSE improvements (included in the 80%)

- **AMD will move to 65nm fabrication only next year**
- **focus on virtualization and security integration**
- **need to catch up in the mobile processor area**
- **currently AMD processors are about 25% more power efficient**

INTEL and AMD offer a wide and large variety of processor types
hard to keep track with new code names

Multi core developments

- dual core dual CPU available right now
- quad core dual CPU expected in the beginning of 2007
- 8-core CPU systems are under development , but not expected to come into market before 2009

(http://www.multicore-association.org/)
cope with change in programming paradigm, multi-threading, parallel

Heterogeneous and dedicated multi-core systems

- Cell processor system PowerPC + 8 DSP cores
- Vega 2 from Azul Systems 24/48 cores for Java and .Net
- CSX600 from ClearSpeed (PCI-X, 96 cores, 25 Gflops, 10W)

Rumor : AMD is in negotiations with ClearSpeed to use their processor board
→ **revival of the co-processor !?**

Game machines

Microsoft Xbox 360 (available, ~450 CHF)
PowerPC based, 3 cores (3.2 GHz each), 2 hardware threads per core
512 MB memory
peak performance = ~ 1000 GFLOPS

Sony Playstation 3 (Nov 2006)
Cell processor, PowerPC + 8 DSP cores
512 MB memory
peak performance = ~ 1800 GFLOPS

problem for High Energy physics :

- **Linux on Xbox**
- **Focus is on floating point calculations, graphics manipulation**
- **Limited memory, no upgrades possible**

INTEL P4 3.0 GHz = ~ 12 GFLOPS ATI X1800XT graphics card = ~ 120 GFLOPS
use the GPU as a co-processor, 32 node cluster at Stony Brook
CPU for task parallelism GPU for data parallelism

compiler exists , quite some code already ported
www.gpgpu.org

Market trends

- **The market share of AMD + INTEL in the desktop PC, notebook PC and server are is about 98 % (21% + 77%)**
- **On the desktop the relative share is INTEL = 18% , AMD = 82%**
  (this is the inverse ratio of their respective total revenues)
- **In the notebook area INTEL leads with 63%**
- **The market share in the server market is growing for AMD, 14% currently**

Largest growth capacity is in the notebook (mobile) market

Worldwide Semiconductor Forecast, 1Q06: Top Applications in 2005 and 2010

Applications	2005 Revenue	2005 Share	Applications	2010 Revenue	2010 Share
PCs	44,858	19.1%	PCs	67,121	19.5%
Digital Cellular	34,864	14.8%	Digital Cellular	60,697	17.6%
Servers	8,700	3.7%	Digital Audio Players	14,451	4.2%
RS3	7,781	3.3%	RS3	11,813	3.4%
Disk Drive	7,711	3.3%	TVs	11,696	3.4%
TVs	7,088	3.0%	Disk Drive	10,645	3.1%
Digital Audio Players	5,436	2.3%	Servers	9,884	2.9%
Manufacturing Systems	5,242	2.2%	Other Automotive	6,542	1.9%
Video Game Machines	5,134	2.2%	Monitor, Flat Panel	6,532	1.9%
Other Automotive	4,807	2.0%	Manufacturing Systems	6,921	2.0%
Other	103,257	44.0%	Other	137,785	40.0%
Total	**234,877**	**100.0%**	**Total**	**344,087**	**100.0%**

RS3: Removable solid-state storage

revenues in million $ units

Source: Gartner Dataquest, February 2006
Semiconductor Forecast Worldwide--Forecast Database [SEQS-WW-DB-DATA]

Gartner

4Q05 Market Headlines: Volume and Growth

Worldwide

Total Shipments:
61.1 Million Units

Year-Over-Year Growth:
16.4%

Quarter-Over-Quarter Growth:
13.2%

Platforms

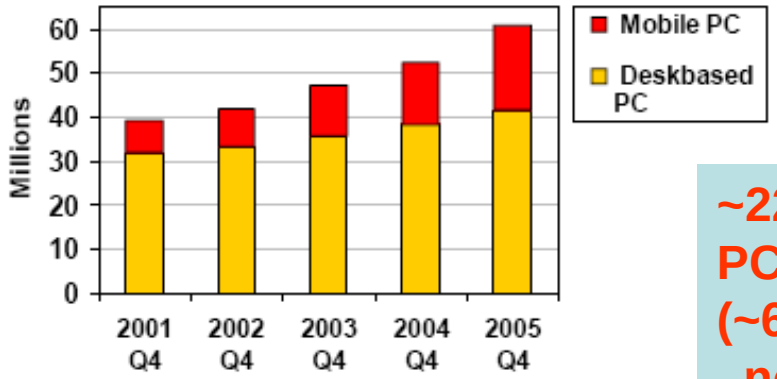
Deskbased PCs
41.6 Million Units

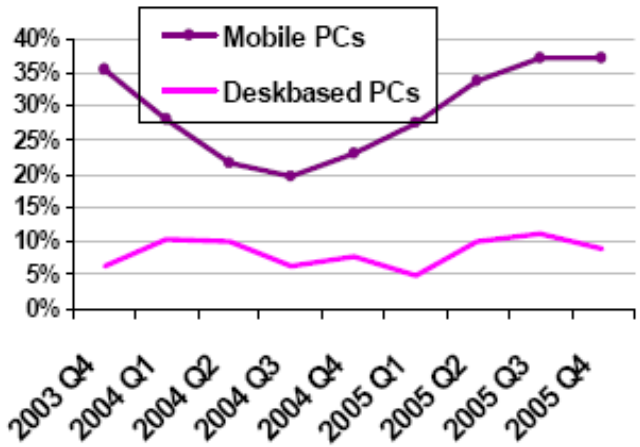
Mobile PCs
19.4 Million Units

Year-Over-Year Growth:

Deskbased PCs:	8.8%
Mobile PCs:	37.1%

Fourth Quarter Comparisons





~220 million PCs in 2005 (~65 million notebooks)

revenue per PC is about 200 $ for the chips

Costs

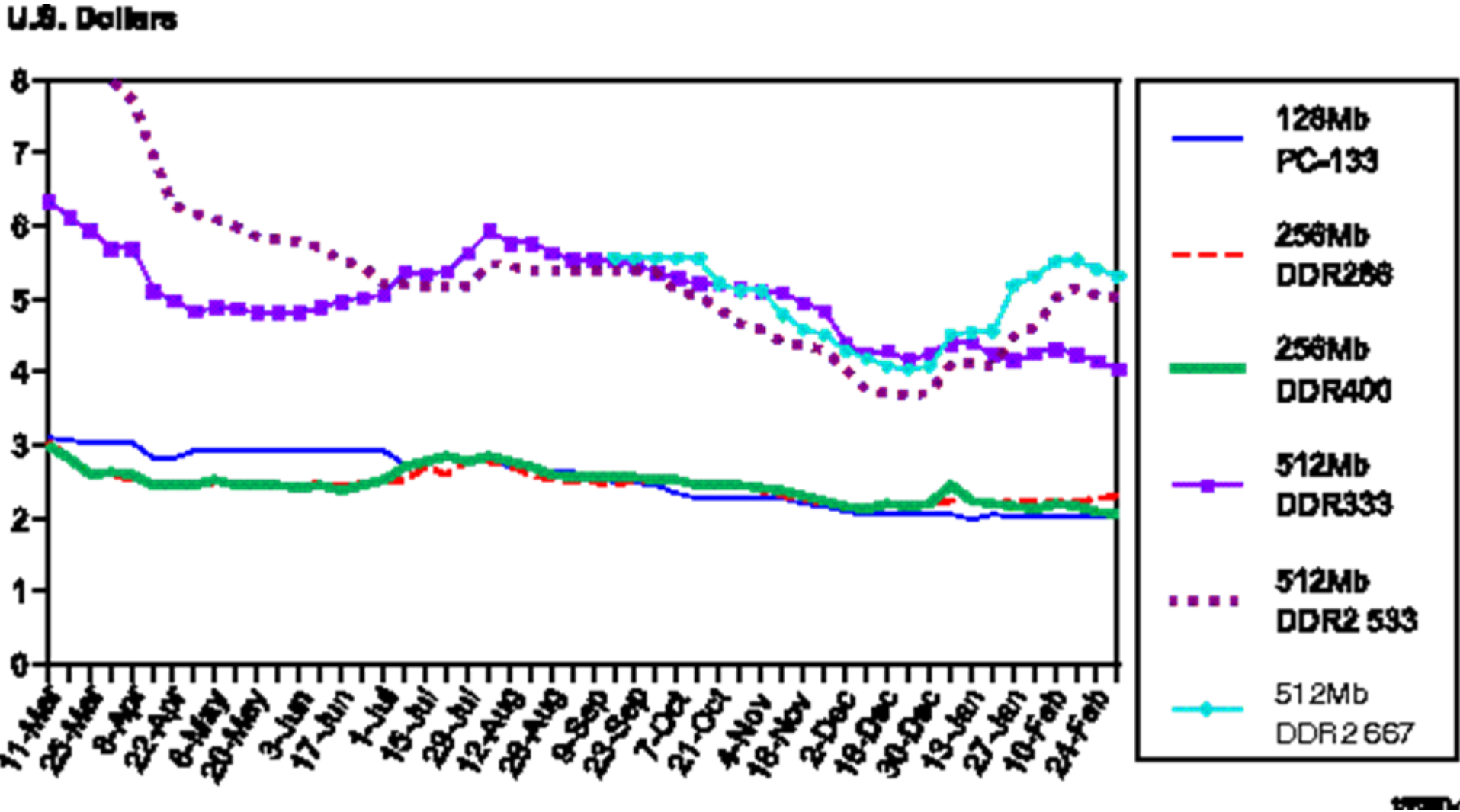
- **More sophisticated motherboards (sound, disk controller, cooling, graphic,etc…)**

- **Major point is memory**
  experiments need 1 (CMS, LHCb) or 2 (ALICE, ATLAS) GByte memory per job
  → multi-core needs corresponding large memory per node,
  4 core == 10 Gbyte memory taking into account to run more jobs per node
  then there are cores to increase CPU efficiency (IO wait-time hiding)
  → per GByte one has to add 10 W power to the system

- **CPU costs include all the extras per core (security-, virtualization-, SSE-, power saving-features, etc.) which we partly can't use.**

dual CPU – single core node cost =	
4 GB memory	32%
2 CPUs	27%
motherboard	20%
chassis+power	16%
hard disk	5%

Spot market , Mbit DIMMs, cost trends



Benchmarks

Problems

- to find ‘your’ configuration under spec.org
- ‘gauge’ the specint values with the programs from the experiments
  → specific programs under steady development
  → needs environment disentanglement for ‘standalone’ programs

Possible solution

- run deep low level processor monitoring (chip level, execution units) continuously in the background on large number of nodes
  → create ‘running’ profiles of the running programs
- compare with running the specint suite on the same nodes

from a talk of Ryszard Jurga, openlab collaboration at CERN

CERN openlab for DataGrid applications

Introduction

- Special on-chip hardware of modern CPU
  - Direct access to CPU resources such as branch prediction, data and instruction caches, floating point instructions, memory operations
  - Event detectors, counters
    - Itanium2: 4 counters, 100+ monitorable events, two set of registers: PMC, PMD
    - **Pentrium4,Xeon**: 44 event detectors, 18 counters
  - Linux interfaces and libraries:
    - Part of kernel in order to per-thread and per-system measurements
    - Perfmon2
      - uniform across all hardware platforms
      - events multiplexing
      - the number of fully supported processors are very low except Itanium
      - kernel 2.6 (integrated for Itanium)
    - **perfctr**

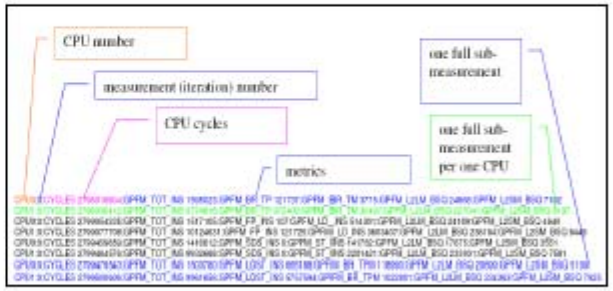
counters exist also on the AMD chips, have not yet looked into that

Monitoring tool - *gpfmon*

- uses perfctr,
- enables multiplexing,
- user and kernel domain,
- per single or total CPU,
- events:

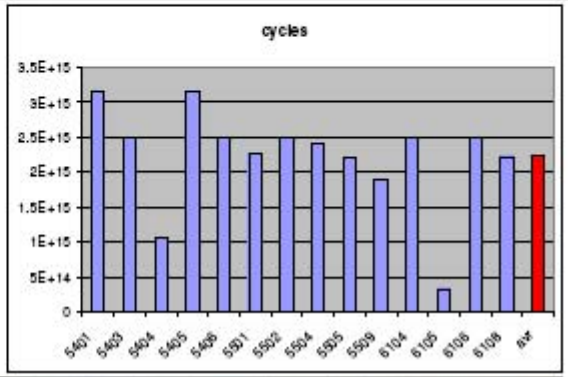
CYC	TOT	BR_TP	BR_TM	L2LM	L2SM
CYC	TOT	FP	LD	L2LM	L2SM
CYC	TOT	SDS	ST	L2LM	L2SM
CYC	TOT	LDST	BR	L2LM	L2SM

CYC – CPU cycles
TOT – Instructions completed
BR_TP – Branch taken predicted
BR_TM – Branch taken mispredicted
L2LM – L2 load missed
L2SM – L2 store missed
FP – Floating point instructions
SDS – scalar instructions
LD – load intstructions
ST – store instructions
BR – BR_TP+BR_TM
LDST - LD+ST



lxbatch monitoring

- 14 machines
- running from 2 day to 2 weeks
- Nocona(10), Irwindale (4)
- 2.8GHz
- 1MB L2(10) 2MB L2(4)
- SL3 (kernel 2.4)



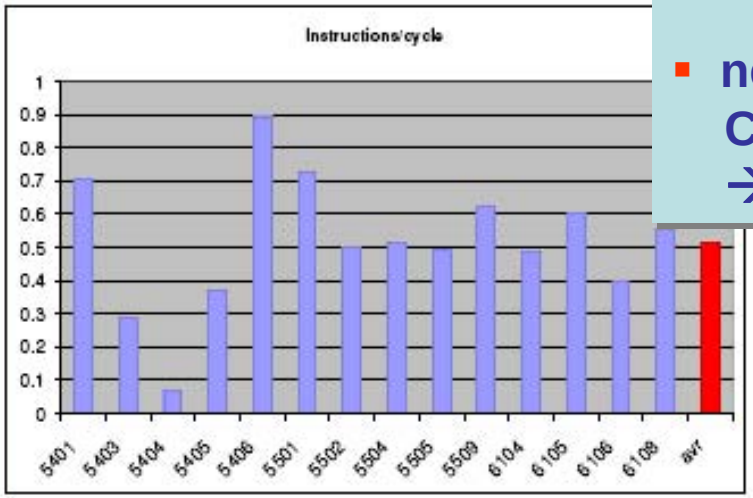
run on different nodes:

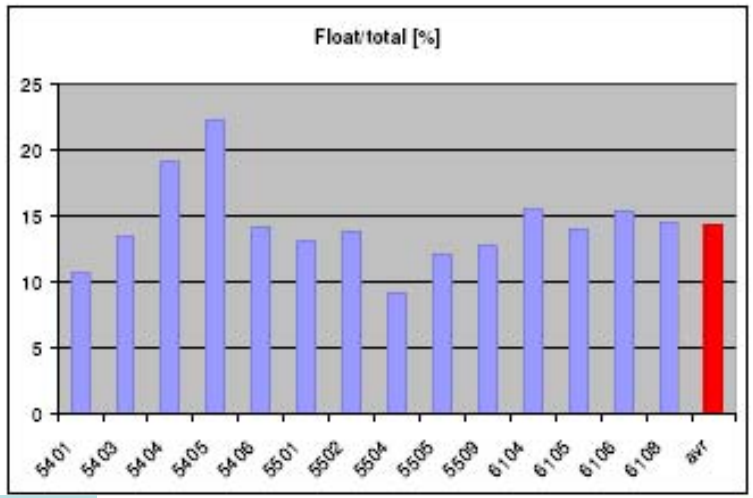
- **CPU architecture**
- **separated by experiments, even only one job per node**
- **mixture of experiments**

CERN openlab
for DataGrid applications

measurements so far show :

- **floating point operations are at the 10-15% level**
- **branch miss-prediction is at the 0.2% level**
- **L2 cache misses are < 0.1 %**
  → deduce memory access performance
  (~ 300 MB/s, needs further investigation..)
- **need 2 cycles per instruction,**
  CPUs can do several instructions per cycle
  → little 'parallelism' in the code





needs more statistics and analysis……