SUPERVISOR: Professor Jukka K. Nurminen

ADVISOR: Zhonghong Ou (POST-DOC.)

Gonçalo Pestana

Energy Efficiency in High Throughput Computing

Tools, techniques and experiments

MOTIVATION: HTC, CERN & energy consumption: Higgs boson and the future

TOOLS: Measuring energy consumption

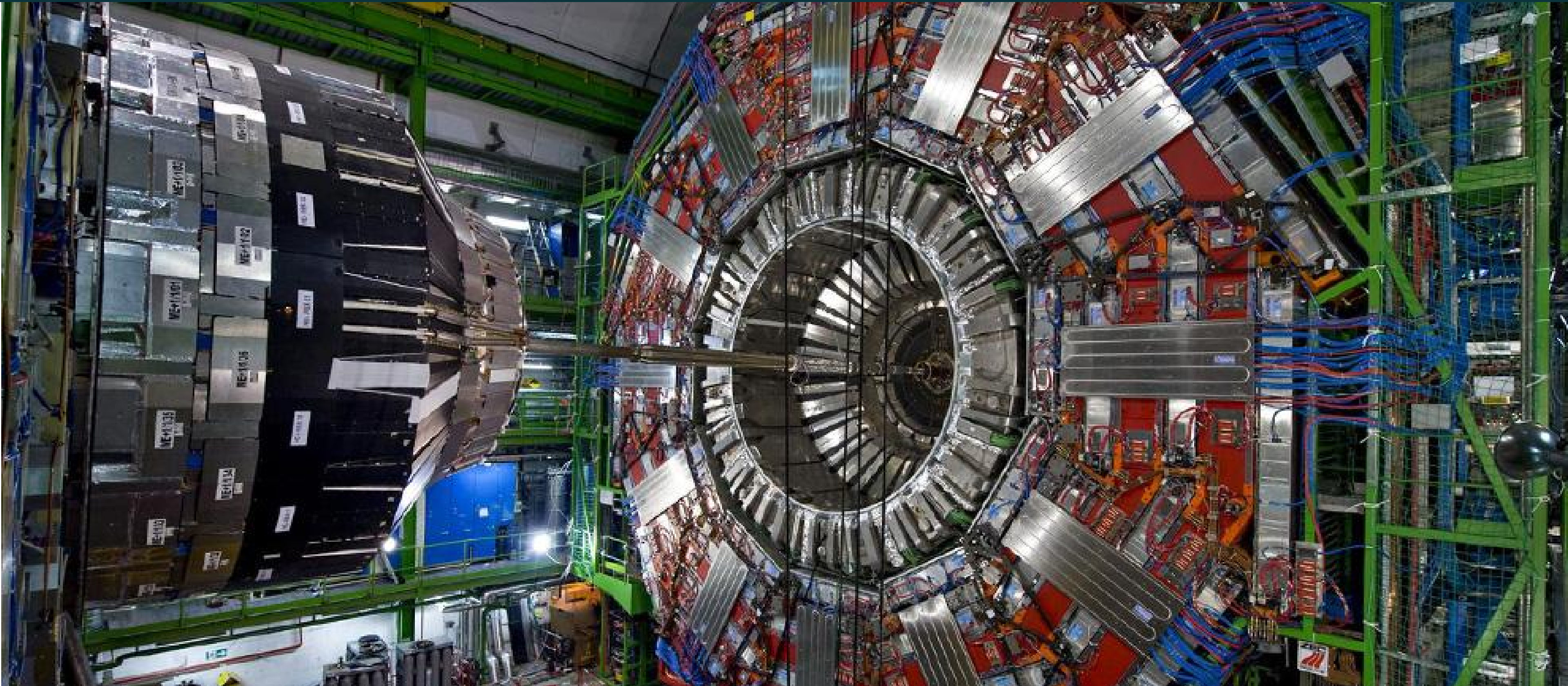
SOLUTION 1: ARM in HTC

SOLUTION 2: Task scheduling algorithm in dynamic pricing markets

CMS collider

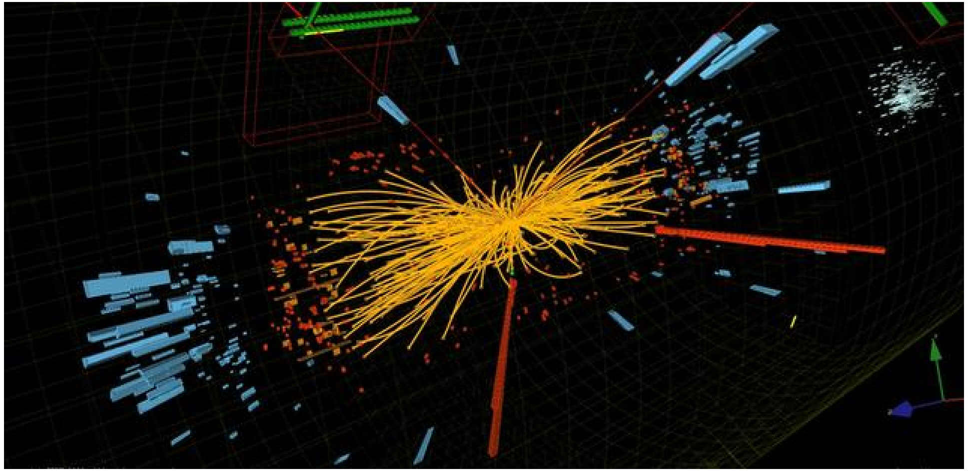
ENERGY EFFICIENCY IN HIGH THROUGHPUT COMPUTING (HTC)

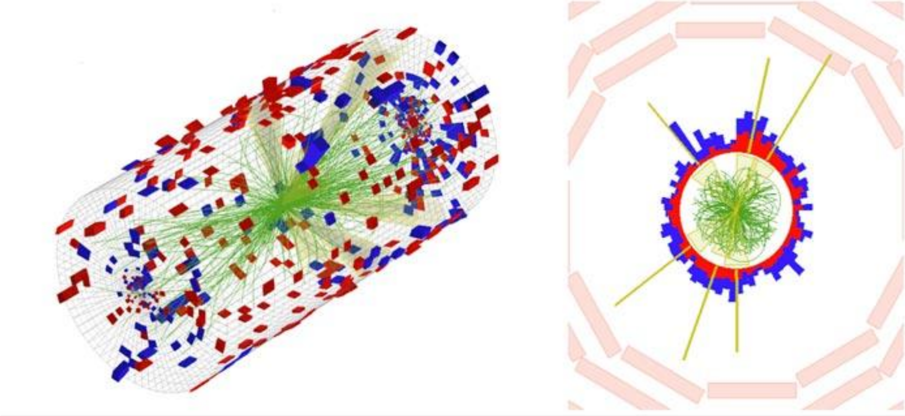
MOTIVATION: HTC, CERN & ENERGY CONSUMPTION



MOTIVATION: HTC, CERN & ENERGY CONSUMPTION

Lots of **data** (1 Petabyte/s → 200 MB/s)





LHC computing power in 2012

MOTIVATION: HTC, CERN & ENERGY CONSUMPTION

In 2012, the Worldwide LHC computing grid *equivalent capacity* of

80,000 to **100,000** x86-64 cores

result: Higgs Boson tracked down

Future

data will increase **2** - **3** orders of magnitude

processing power in proportion

Expectable to happen throughout all HTC industry

question

How to decrease electricity bill ?

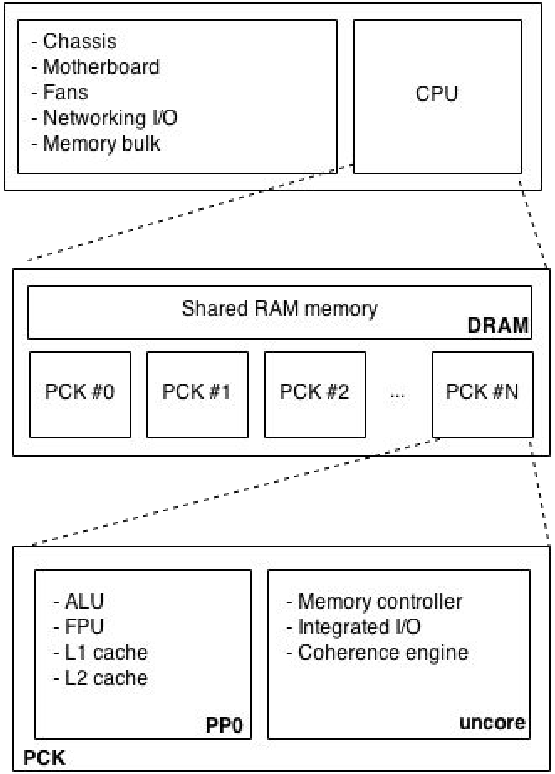
first: measure

TOOLS: MEASURING ENERGY EFFICIENCY

Techniques and **tools** for measuring power consumption are important ...

TOOLS: MEASURING ENERGY EFFICIENCY

… and systems are **complex**

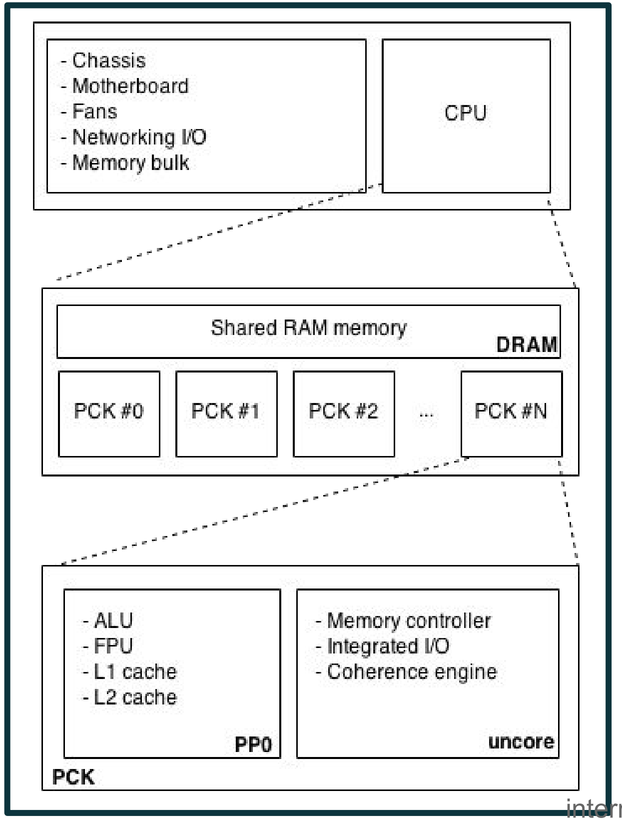


several layers and granularities

TOOLS: MEASURING ENERGY EFFICIENCY

External measurements

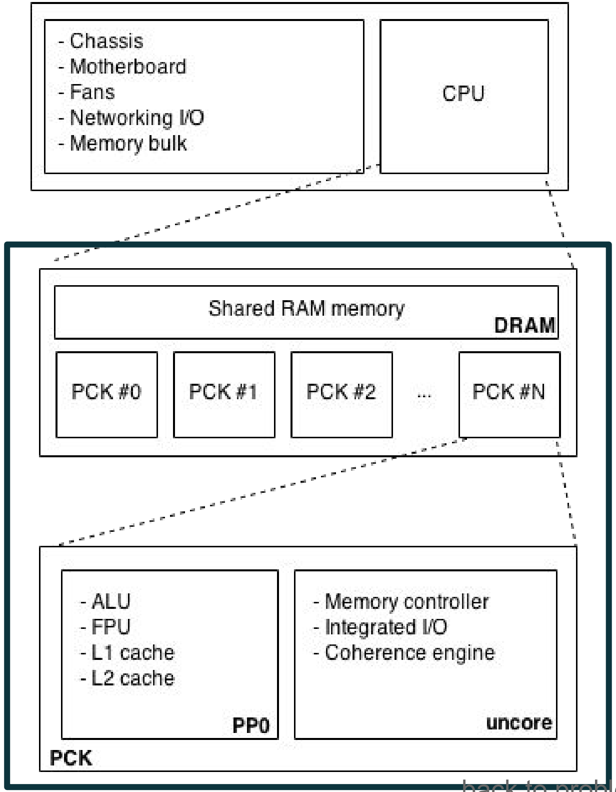
Power consumed without breaking down the system into components



TOOLS: MEASURING ENERGY EFFICIENCY

On-chip measurements

Power consumed by different components of the CPU



back to problem

back to our main problem:

How to decrease electricity bill ?

SOLUTION 1: ARM IN HTC?

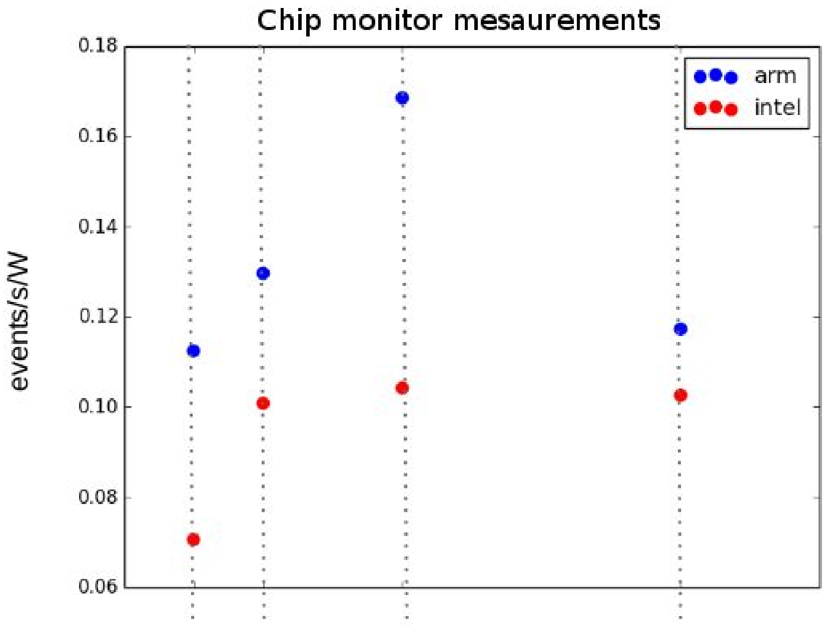
Are smartphones' CPUs

the future of High Throughput Processing ?

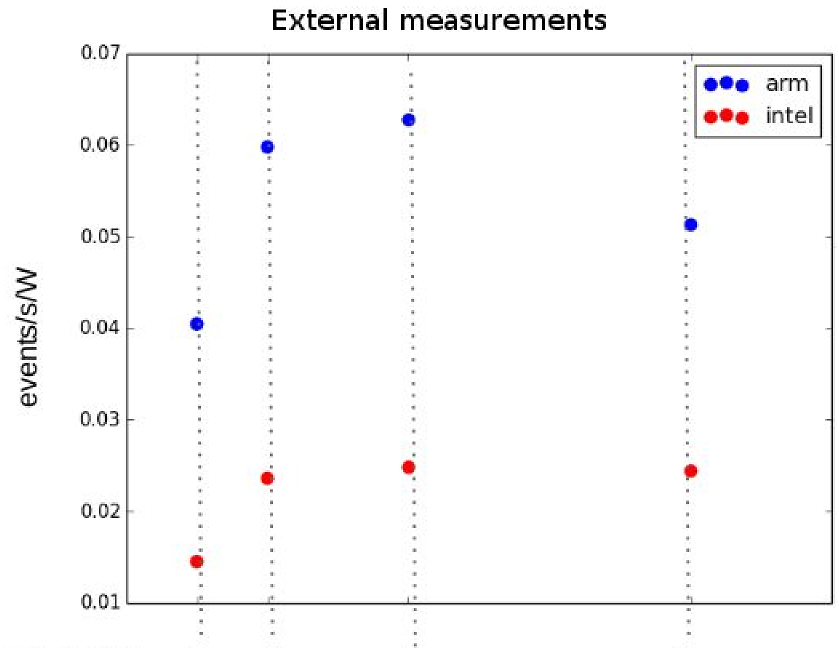
Experiments

ARCHITECTURES:	**ARM** machine *vs* **INTEL** machine
WORKLOAD:	events (ratio nr. events/per core variable)
ENERGY PERF METRIC:	events/second/watts (e/s/W) → the higher the better
MEASUREMENTS:	external and internal

SOLUTION 1: ARM IN HTC?



Number events/core	¼	½	1	2
ARM	1	2	4	8
Intel	8	16	32	64



Number events/core	¼	½	1	2
ARM	1	2	4	8
Intel	8	16	32	64

SOLUTION 1: ARM IN HTC?

energy efficiency

vs

speed

back to problem

back to our main problem:

How to decrease electricity bill ?

SOLUTION 2: DYNAMIC PRICING AND TASK SCHEDULING

Algorithm to schedule tasks across different computing architectures in a dynamic pricing energy market

which simplifying...

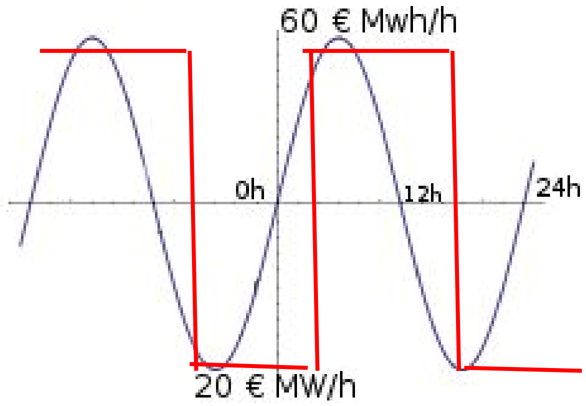
SOLUTION 2: DYNAMIC PRICING AND TASK SCHEDULING

Use expensive machines when energy is cheaper

algorithm assumes that...

SOLUTION 2: DYNAMIC PRICING AND TASK SCHEDULING

Dynamic electricity price



Machines with different energy profiles

ARM more energy efficient, slower

INTEL less energy efficient, faster

SOLUTION 2: DYNAMIC PRICING AND TASK SCHEDULING

Idea *(simplified)*

Schedule tasks to **INTEL** when electricity is cheaper

Schedule tasks to **ARM** when electricity is more expensive

SOLUTION 2: DYNAMIC PRICING AND TASK SCHEDULING

How

Algorithm that computes which machines should compute data based on:

- **deadline** (how many tasks in how much time)
- **energy pricing dynamics**
- **energy profiling of the machines**

CONCLUSIONS

PROBLEM: Energy consumption is bottleneck in HTC

SOLUTION 1: ARM shows potential for HTC (but tradeoffs)

SOLUTION 2: Scheduling tasks based on energy pricing

How to measure energy consumption in complex systems ?

thanks, questions?

SUPERVISOR: Professor Jukka K. Nurminen
ADVISOR: Zhonghong Ou (POST-DOC.)

Gonçalo Pestana

Energy Efficiency in High Throughput Computing

Tools, techniques and experiments

Q & A

backup

IgProf

IgProf

application profiler developed at CERN by the CMS software team

general purpose. open source. not experiment specific

measures performance (time spent in functions) and memory usage at *runtime*

allows developer to understand:

- bottlenecks
- where code needs to be optimised

cross platform: recently ported to 64-bit ARM, also supports 32-bit ARM, Intel x86 and x86-64

IgProf & energy profiling

Uses RAPL and PAPI to measure energy consumed.

Map functions and low level operations with **energy consumption**

more info (strategies, results, examples)

paper *and* ***http://igprof.org/***

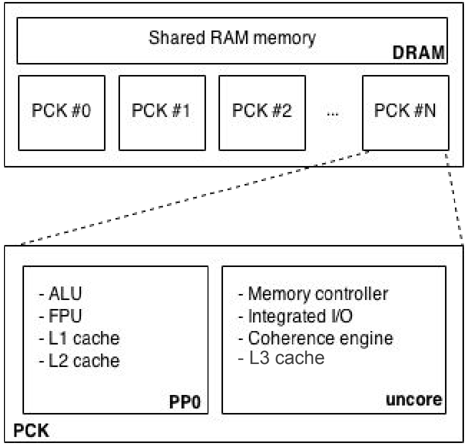
RAPL

Running Average Power Limiting *(RAPL) by Intel*

Provides a platform for power monitoring and power limiting of SoC.

Different sampling **domains**

package *(PKG)*, DRAM, core



Low level measurements package, cores, dram

Resolution according to Intel, sampling frequency up to ~1 kHz

Power capping is also supported by RAPL

Accuracy high (according to *Intel*)

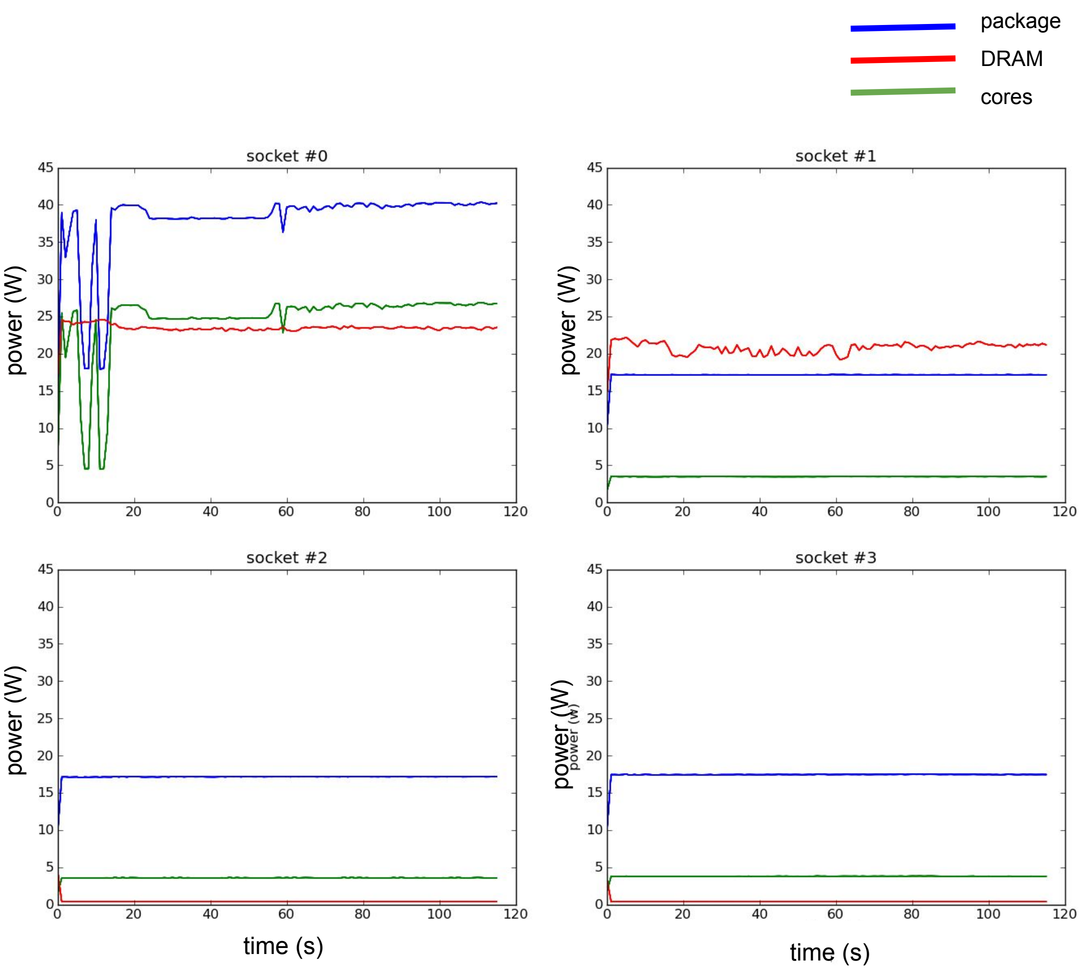
Example of RAPL

System with 4 sockets

Sockets #0 and #1 working

Sockets #2 and #3 idle

Possible to understand how **packages**, **cores** & **dram** consume energy



Comparison ARM vs Intel

Comparison ARMv7 vs Intel XEON

32bit ARMv7 is used on smartphones. comparison with Intel XEON

measurements

Internal

RAPL for *Intel*

cross platform *chip monitor* integrated (TI INA 231) for ARMv7

External

workload

ParFullCMS

Multithreaded

Geant4 benchmark application

Uses the CMS geometry

ARMv7	Exynos5 Octa Cortex™
	4x A15 @ 1.6Ghz and/or A7 cores (big.LITTLE technology)
	2 GB RAM
	ARMv7/32bit
	development board
Intel	**32x** Intel™ Xeon™ CPU E5-2650 @ 2.00GHz
	252 GB RAM
	system on a rack

Dynamic Pricing algorithm

Dynamic Pricing *(considering power in ev/s/W conversion)*

ARM

workload 0.063 ev/s/W * 7.5 W = 0.4725 ev/s (40 824 ev/day)

price 0.0003 euros per hour per machine (0.0072 euros/day)

(0.0072 euros/day) / (40 824 ev/day) = **1.7*10^-7 euros/ev**

INTEL

workload 0.023 ev/s/W * 368 W = 8.464 ev/s (731 289 ev/day)

price 0.01472 euros per hour per machine (0.35328 euros/day)

(0.35328 euros/day) / (731 289 ev/day) = **4.8*10^-7 euros/ev**

Using algorithm

Data needed:

	Price per event	Total events per day
bucket 1 (ARM low)	8.8 *10^-8 €	40 824/2
bucket 2 (ARM high)	2.6 *10^-7 €	40 824/2
bucket 3 (Intel low)	2.4 *10^-6 €	731 289/2
bucket 4 (Intel high)	7.2 *10^-6 €	731 289/2

Output:

Final price

		Events processed					
		1 day	2 days	10 days	30 days	75 days	100 days
ARM	**bucket_1** (20€/Mwh)	*not possible*	40824	204120	612360	1200000	1200000
	bucket_2 (60€/Mwh)	*not possible*	40824	204120	587640	0	0
INTEL	**bucket_3** (20€/Mwh)	*not possible*	731288	791760	0	0	0
	bucket_4 (60€/Mwh)	*not possible*	387064	0	0	0	0
Total Price		-	4.55€	1.97€	0.206€	0.106€	0.106€

Two machines. 120 000 events. variable deadline

Scheduling Algorithm

```
days = 10
nr_events = 1200000

bucket_1 = {
    'price_ev': 8.8*10**-8,
    'nr_ev_day': 40824/2,
    'name': 'ARM_LOW'
}

bucket_2 = {
    'price_ev': 2.6*10**-7,
    'nr_ev_day': 40824/2,
    'name': 'ARM_HIGH'
}

bucket_3 = {
    'price_ev': 2.4*10**-6,
    'nr_ev_day': 731289/2,
    'name': 'INTEL_LOW'
}

bucket_4 = {
    'price_ev': 7.2*10**-6,
    'nr_ev_day': 731289/2,
    'name': 'INTEL_HIGH'
}

result = scheduler([bucket_1, bucket_2, bucket_3, bucket_4], days, nr_events)
print result
```

Scheduling Algorithm

```python
#!/usr/bin/python
def scheduler(buckets, days, nr_events):
    nr_events_left = nr_events
    final_prices = []
    sorted(buckets) #sorts according to price per event, from lowest to highest

    for bucket in buckets:
        nr_possible_ev_process = bucket['nr_ev_day'] * days

        #ensures that no more events than the needed are processed
        nr_ev_process = nr_events_left if nr_possible_ev_process > nr_events_left \
            else nr_possible_ev_process

        price = bucket['price_ev'] * nr_ev_process
        final_prices.append(price)
        print bucket['name']+' processed '+ str(nr_ev_process)

        nr_events_left -= nr_ev_process
        if (nr_events_left <= 0):
            return sum(final_prices)

    return 'ERR: not enough machine processing power to process events before the deadline'
```

Scheduling Algorithm (comparison/ inspiration)

Jobshop algorithm

→ limited set of tasks

→ limited set of resources

→ tries to schedule the tasks to the machines so that all tasks are completed the least amount of time possible

Differences:

→ **limited set of time**

→ **machines are not equal** (different energy profile)

→ **Not online algorithm**: (*online: each job is presented and algorithm has to make decision before next job*)

In our algorithm, all the information the algorithm needs to schedule the tasks is given upfront (for a defined chunk of time);

Others

ARMv8 in HTC

Under the circumstances of the experiment, the overall results show that APM X-Gene is 2.73 slower than Intel Xeon Phi. From the energy consumption performance (events per second per watt), the Intel Xeon E-2650 is the most efficient, with APM X-Gene presenting similar performance despite the **absence of platform specific optimizations.** Therefore, \cite{ACAT14ARMDAVID} concludes by stating that the APM X-Gene 1 Server-On-Chip ARMv8 64-bit solution is relevant and potentially interesting platform for heterogeneous high-density computing.

source: Abdurachmanov, D., Bockelman, B., Elmer, P., Eulisse, G., Knight, R., and Muzaffar, S. ***Heterogeneous high throughput scientific computing with APM x-gene and intel xeon phi***. CoRR abs/1410.3441 (2014).