MHPC
Master in High Performance Computing

FHPC course:
the energy issues on HPC system.

Stefano Cozzini

CNR/IOM and eXact-lab srl

SISSA
Scuola Internazionale Superiore di Studi Avanzati

ICTP The Abdus Salam International Centre for Theoretical Physics

MHPC
Master in High Performance Computing

Agenda

- Energy problem in HPC
- Green500
- Available technology on modern CPUs
- How can we measure it ?

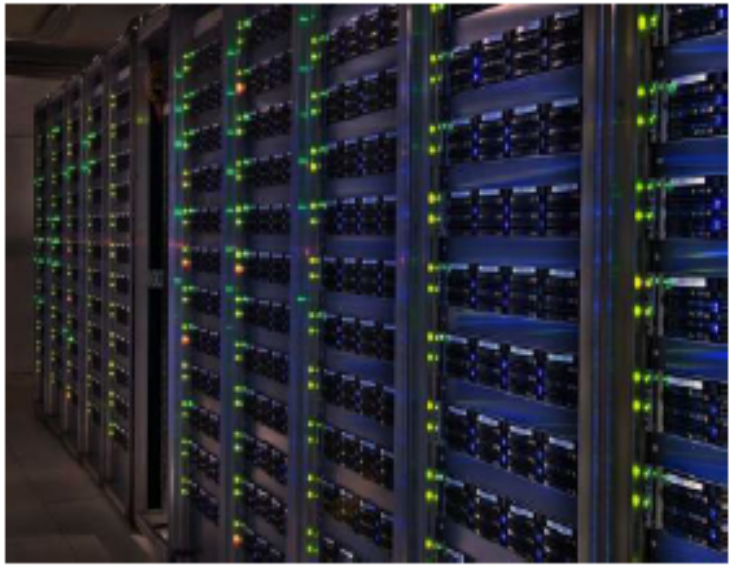
- AIM:
- Give you the feeling how much is important the Energy problem in the HPC arena right now

SISSA ICTP

MHPC
Master in High Performance Computing

WE ARE NOT GREEN

A NATIONAL SUPERCOMPUTING FACILITY HAS A YEARLY CO_2 FOOTPRINT COMPARABLE TO A TAKE OFF OF A SATURN V



=

SISSA

MHPC
Master in High Performance Computing

WE ARE NOT GREEN

A simulation of 500 ns with 90 k atoms = 1week on 512 cores = 3200 kWh

Correspond to:

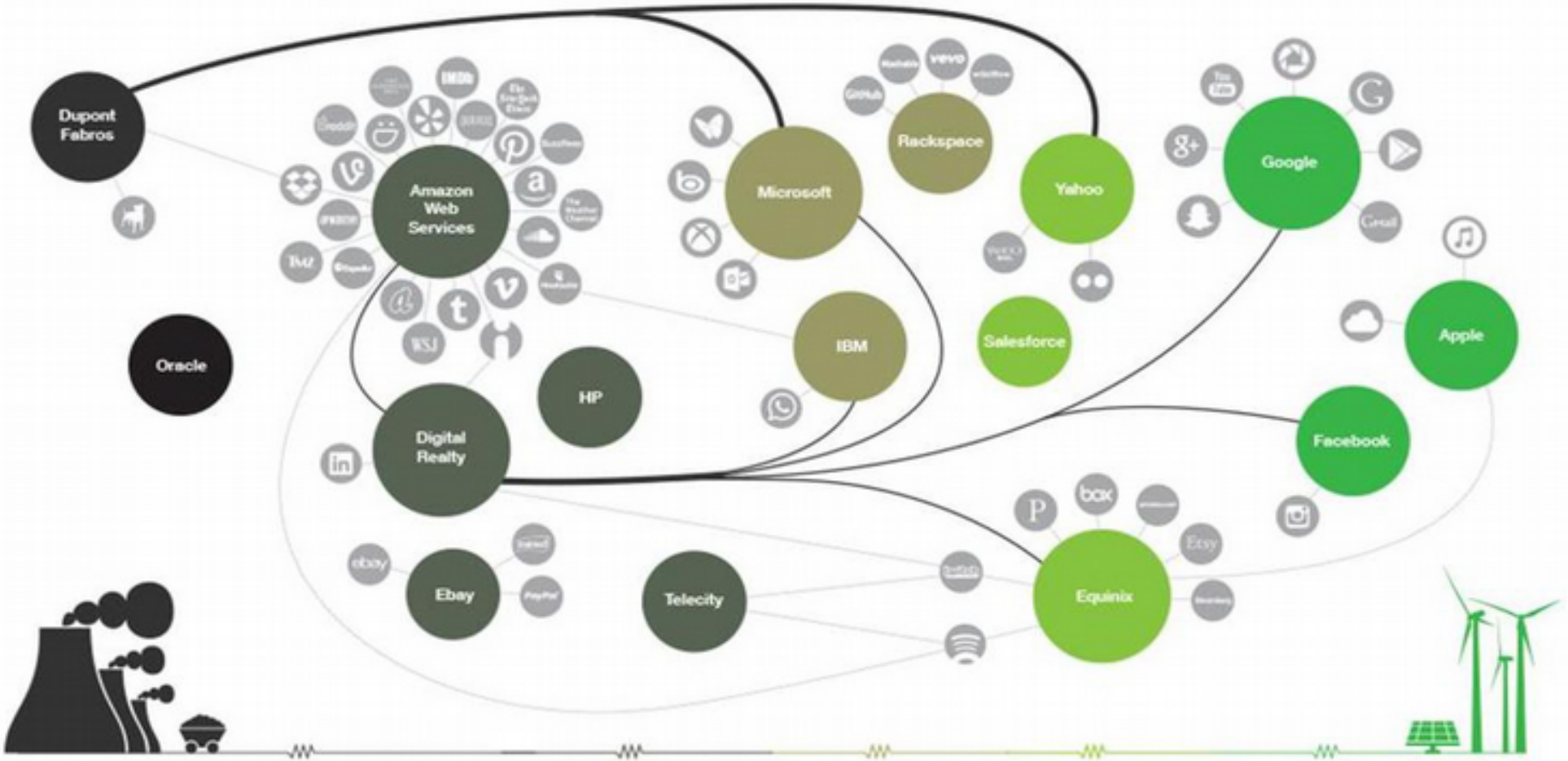
a) 1600 CO_2 kg

b) 340 € energy bill

c) 13000 km by car

MHPC
Master in High Performance Computing

Someone are darks and someone green (at least they claim to be..)



MHPC
Master in High Performance Computing

Top Exascale Challenges

- System Power & Energy
- System efficiency & cost
- New, efficient, memory subsystem
- Extreme parallelism, data locality
- Resiliency to provide system reliability

SISSA ICTP

MHPC
Master in High Performance Computing

Road to exascale (from 2008 paper)

- Goal

  "Because of the difficulty of achieving such physical constraints, the study was permitted to assume some growth, perhaps a factor of 2X, to something with a maximum limit of 500 racks and 20 MW for the computational part of the 2015 system."

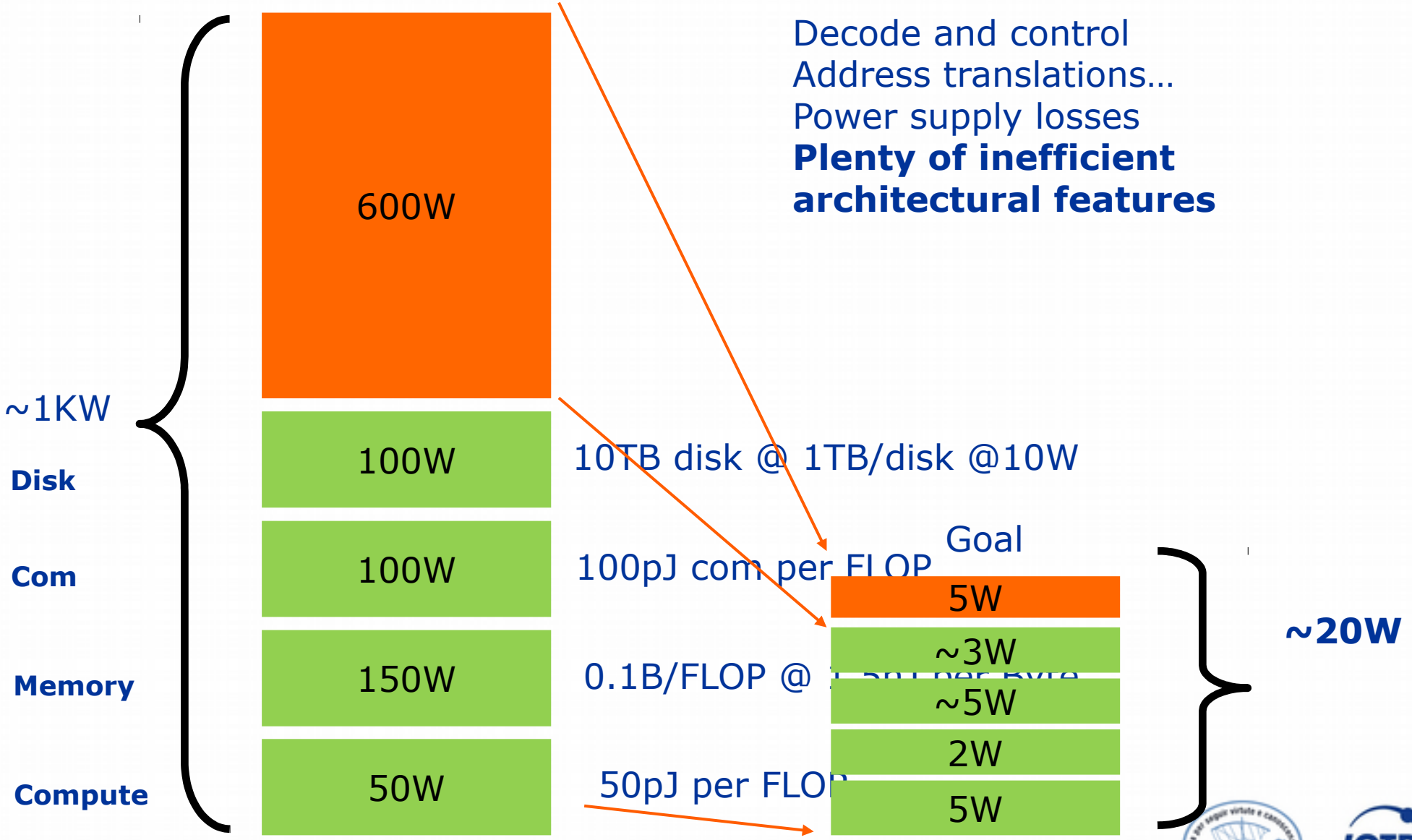
- Realistic Projection?

  "Assuming that Linpack performance will continue to be of at least passing significance to real Exascale applications, and that technology advances in fact proceed as they did in the last decade (both of which have been shown here to be of dubious validity), then […] an Exaflop per second system is possible at around 67 MW."

SISSA ICTP

Where is the Energy Consumed?

Teraflop system today



Decode and control
Address translations…
Power supply losses
Plenty of inefficient architectural features

MHPC
Master in High Performance Computing

The power wall

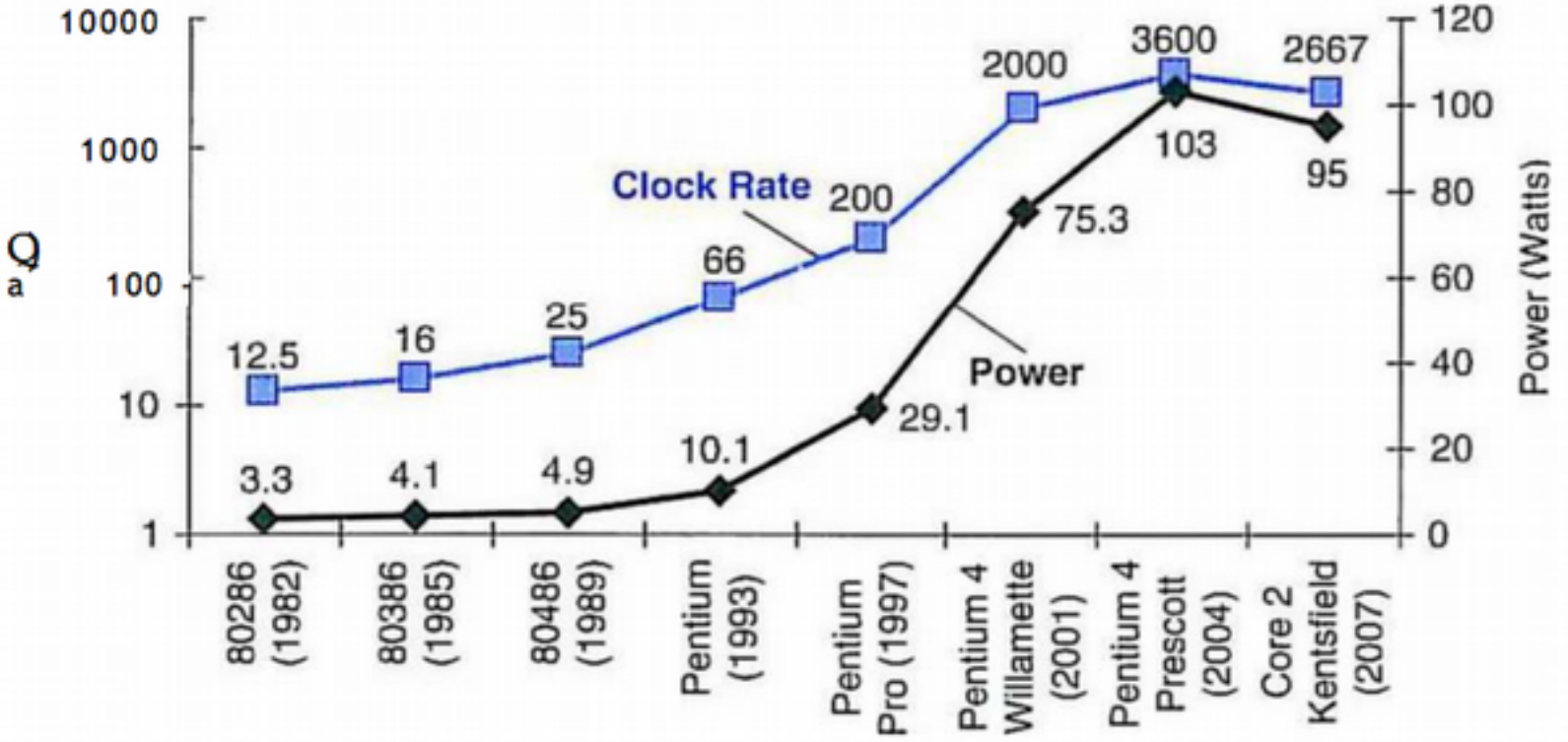


FIGURE 1.15 Clock rate and Power for Intel x86 microprocessors over eight generations and 25 years. The Pentium 4 made a dramatic jump in clock rate and power but less so in performance. The Prescott thermal problems led to the abandonment of the Pentium 4 line. The Core 2 line reverts to a simpler pipeline with lower clock rates and multiple processors per chip.

ICTP

MHPC
Master in High Performance Computing

Where do we need energy in CPUs ?

- CMOS, the primary source of power dissipation is so-called dynamic power —that is, power that is consumed during switching.

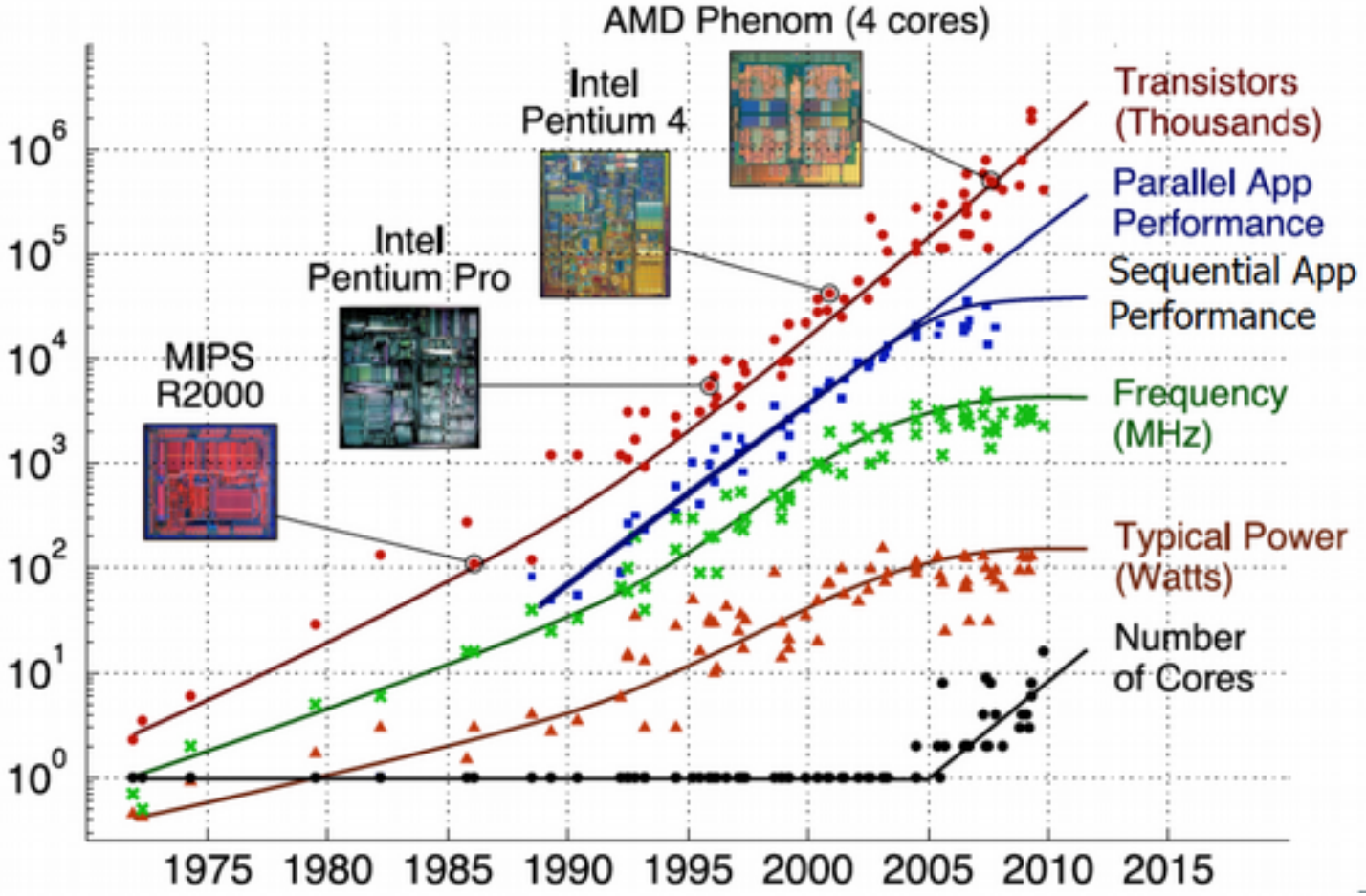
  Power = Capacitive load x $(\text{Voltage})^2$ x Frequency switched

- How could clock rates grow by a factor of 1000 while power grew by only a factor of 30? Power can be reduced by lowering the voltage, which occurred with each new generation of technology, and power is a function of the voltage squared.
- Typically, the voltage was reduced about 15% per generation. In 20 years, voltages have gone from 5V to 1 V, which is why the increase in power is only 30 times
- We reached that limit a few year ago: not possible to decrease any further the Voltage.

SISSA ICTP

MHPC
Master in High Performance Computing

Transition to multicore...



SISSA

ICTP

MHPC
Master in High Performance Computing

The Ultimate Goal of "The Green500 List"

- **Raise awareness of energy efficiency in supercomputing.**
  - Drive energy efficiency as a first-order design constraint (on par with FLOPS).

Encourage fair use of the list rankings to promote energy efficiency in high-performance computing systems.

THE GREEN 500

MHPC
Master in High Performance Computing

Green500

Green500 List: the 500greenest supercomputers (from 2007)

energy efficiency of supercomputer = performance per watt" (PPW)

PPW = Performance / Power

- In Green500:
  - Performance: *the achieved maximal performance by the Linpack benchmark on the entire system, denoted as Rmax*
  - Power: *average system power consumption during the execution of Linpack with a problem size that delivers Rmax*

SISSA ICTP

MHPC
Master in High Performance Computing

Green500 (june 2017)

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	259	**Shoubu system B** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , **PEZY Computing / Exascaler Inc.** Advanced Center for Computing and Communication, RIKEN Japan	794,400	842.0	50	17.009
2	307	**Suiren2** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , **PEZY Computing / Exascaler Inc.** High Energy Accelerator Research Organization /KEK Japan	762,624	788.2	47	16.759
3	276	**Sakura** - ZettaScaler-2.2, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband EDR, PEZY-SC2 , **PEZY Computing / Exascaler Inc.** PEZY Computing K.K. Japan	794,400	824.7	50	16.657
4	149	**DGX SaturnV Volta** - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , **Nvidia** NVIDIA Corporation United States	22,440	1,070.0	97	15.113
5	4	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , **ExaScaler** Japan Agency for Marine-Earth Science and Technology Japan	19,860,000	19,135.8	1,350	14.173

MHPC
Master in High Performance Computing

Green500 (june 2018)

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	359	**Shoubu system B** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. Advanced Center for Computing and Communication, RIKEN Japan	794,400	857.6	47	18.404
2	419	**Suiren2** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. High Energy Accelerator Research Organization /KEK Japan	762,624	798.0	47	16.835
3	385	**Sakura** - ZettaScaler-2.2, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. PEZY Computing K.K. Japan	794,400	824.7	50	16.657
4	227	**DGX SaturnV Volta** - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , Nvidia NVIDIA Corporation United States	22,440	1,070.0	97	15.113
5	1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,282,544	122,300.0	8,806	13.889

MHPC
Master in High Performance Computing

Moving data IS the problems.

Figure 5: Bandwidth vs system distance

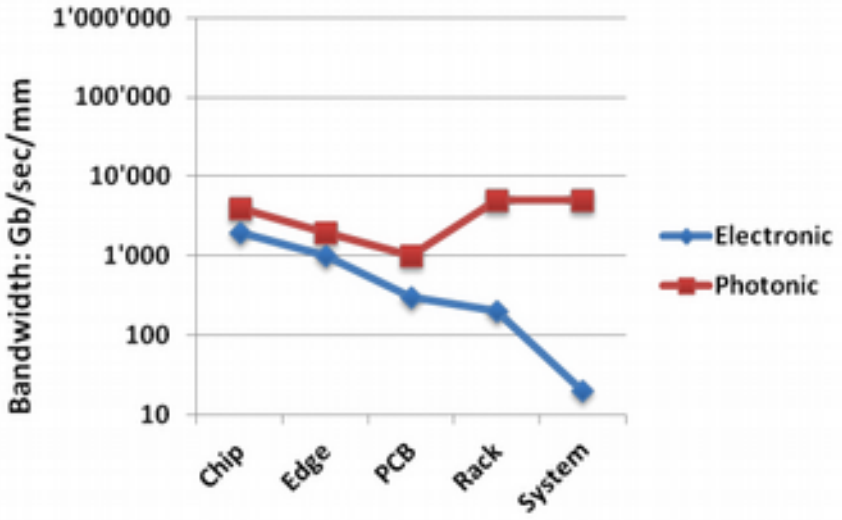
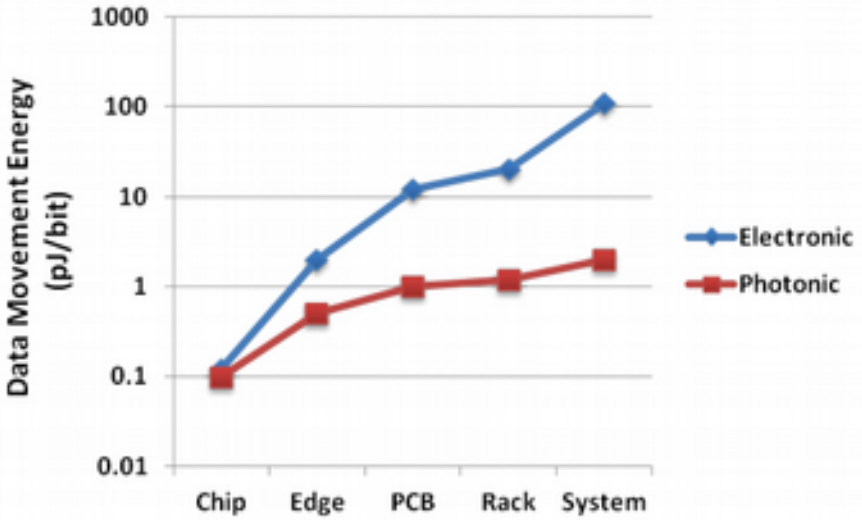


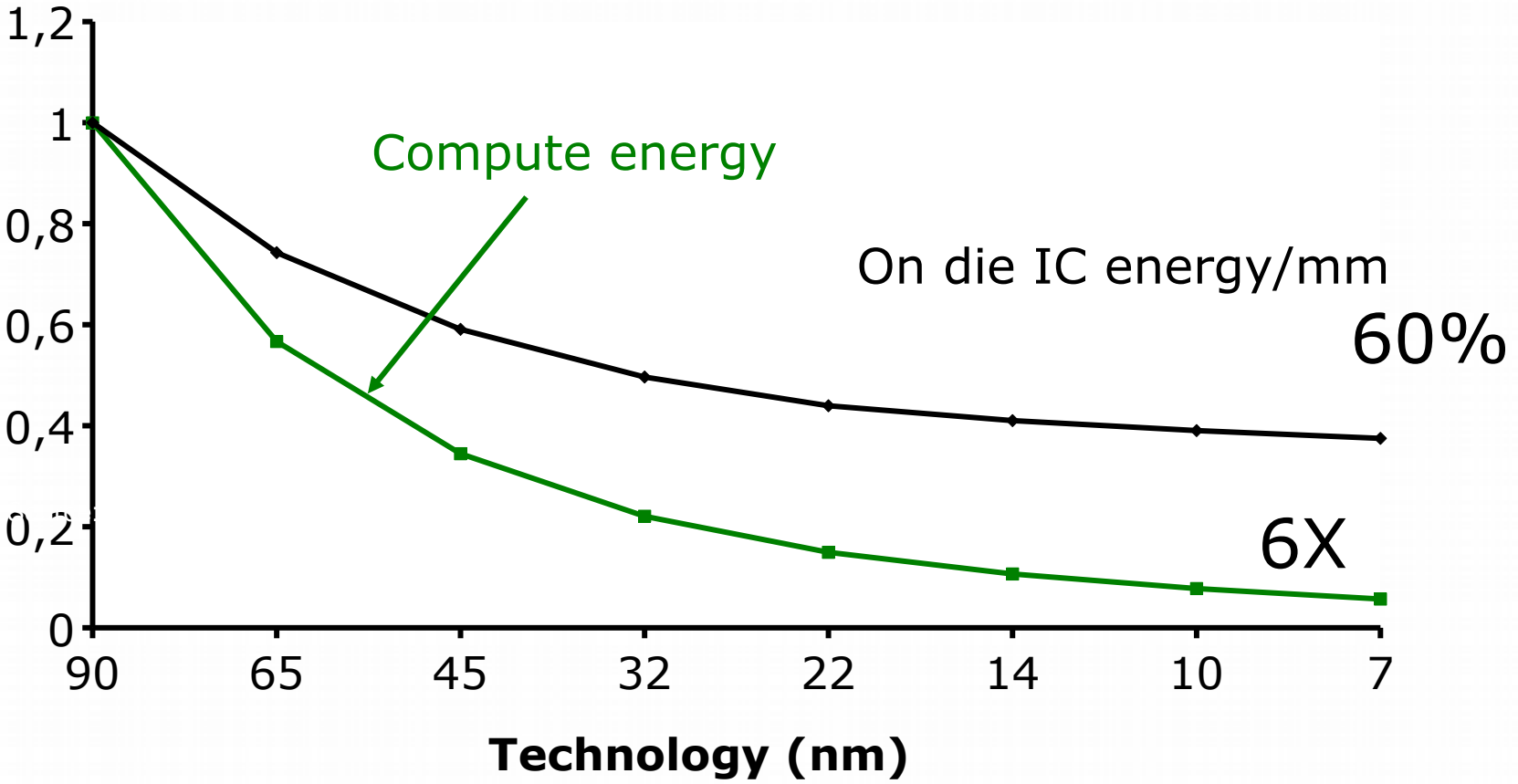
Figure 6: Energy vs system distance



From: The Top Ten Exascale Research Challenges (2014) available at: http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20140210/Top10reportFEB14.pdf

SISSA ICTP

On-die Data Movement vs Compute



Interconnect energy (per mm) reduces slower than compute
On-die data movement energy will start to dominate

MHPC
Master in High Performance Computing

How to reduce the power consumption of HPC resources?

- policy-based automatic power management
  (idle nodes into power saving modes, power on/wake nodes for new workload, ...)
- exploit hardware capabilities: DVFS / power-saving states / performance states / turbo mode
- power capping policies (maximum amount of overall admitted power consumption)
- assign workload to highest performance-per-watt resources first
- energy-aware resource management systems and schedulers able to exploit all of the above, implementing out-of-band and unattended energy assessment capabilities

SISSA ICTP

MHPC
Master in High Performance Computing

Hardware capabilities

- Advanced Configuration and Power Interface (ACPI) defines:
  - sleeping states (S-states)
  - power (core) states (C-states)
  - performance states (P-states)
- Such of the above solutions involves methods like
  - DVFS (Dynamic Voltage and Frequency Scaling)
  - RFTS(Run Fast Then Stop) mechanisms and power/clock gating
  - Turbo Boost technology (INTEL specific)
- The above tricks dynamically configure and monitor the power consumption
- All these features, which are implemented at hardware level by the CPUs, can be enabled by compliant motherboard's BIOS and exposed as a control knob to the operating system for run-time power-optimization.

SISSA ICTP

MHPC
Master in High Performance Computing

Some more detailed definitions

- Thermal Design Power (TDP):
  - is the maximum amount of heat generated by a computer chip or component (often a CPU, GPU or system on a chip) that the cooling system in a computer is designed to dissipate under any workload.
  - The processor's rated frequency assumes that all execution cores are running an application at TDP level
- P-states:
  - During the execution of code, the operating system and CPU can optimize power consumption through different p-states (performance states). Depending on the requirements, a CPU is operated at different frequencies. P0 is the highest frequency (with the highest voltage).
- C-states:
  - nlike the P-States, which are designed to optimize power consumption during code execution, C-States are used to optimize or reduce power consumption in idle mode (i. e. when no code is executed).

SISSA ICTP

MHPC
Master in High Performance Computing

HW tricks: DVFS

- Observation
  - Power α voltage^2 α frequency
  - Performance α frequency
- Mechanism: Dynamic Voltage & Frequency Scaling (DVFS)
  - Allows changes to CPU voltage & frequency at run time
  - Trades CPU performance for power reduction
- Uses commodity technology
- Policy: DVFS Scheduling
- Determines
  - WHEN to adjust the frequency-voltage setting, and
  - WHAT the new setting should be.

SISSA ICTP

Master in High Performance Computing

How can I change the frequency ?

```
How to decide what frequency within the CPUfreq policy should be used?
That's done using "cpufreq governors".

Basically, it's the following flow graph:

CPU can be set to switch independently    |         CPU can only be set
      within specific "limits"            |       to specific frequencies

                             "CPUfreq policy"
                consists of frequency limits (policy->{min,max})
                     and CPUfreq governor to be used
                         /                    \
                        /                      \
                       /                       the cpufreq governor decides
                      /                        (dynamically or statically)
                     /                         what target_freq to set within
                    /                          the limits of policy->{min,max}
                   /                                \
                  /                                  \
        Using the ->setpolicy call,              Using the ->target/target_index/fast_switch call,
            the limits and the                    the frequency closest
             "policy" is set.                     to target_freq is set.
                                                  It is assured that it
                                                  is within policy->{min,max}
```

From: https://www.kernel.org/doc/Documentation/cpu-freq/governors.txt

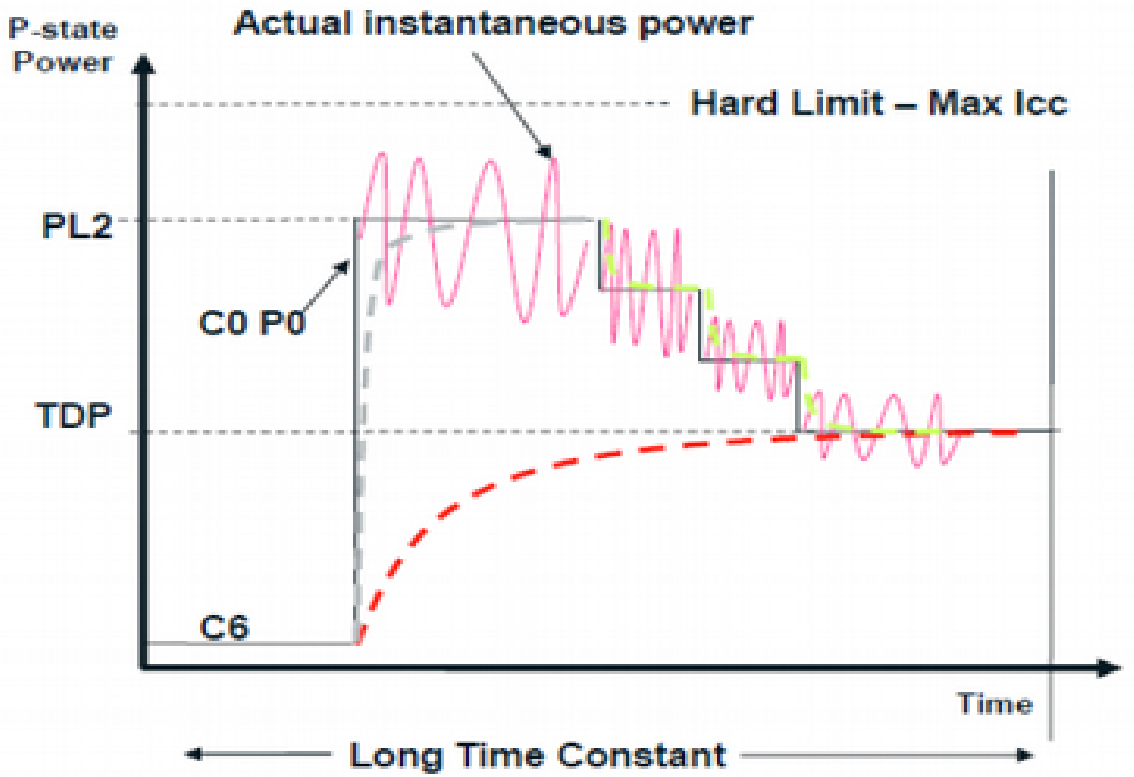
MHPC
Master in High Performance Computing

HW tricks: Turbo boost mode..

Intel® Turbo Boost Technology Behavior

Haswell Intel® Turbo Boost Technology uses transient headroom:

- Can briefly exceed TDP for maximum performance.
- Temperatures ramp more quickly, but no impact to "steady state" condition.
- Transient power limited by power delivery capacities of platform (PL2).



MHPC
Master in High Performance Computing

How to measure the power consumption?

- **external** hardware-dependent probes and sensors, PDUs, power grid counters
- **local** machine hardware sensors
- metrics obtained accessing hardware counters available on most major **microprocessors**

  RAPL (Running Average Power Limit) (MSR by INTEL)
  - monitors, controls, and gets notifications on SoC power consumption (platform level power capping, monitoring and thermal management)

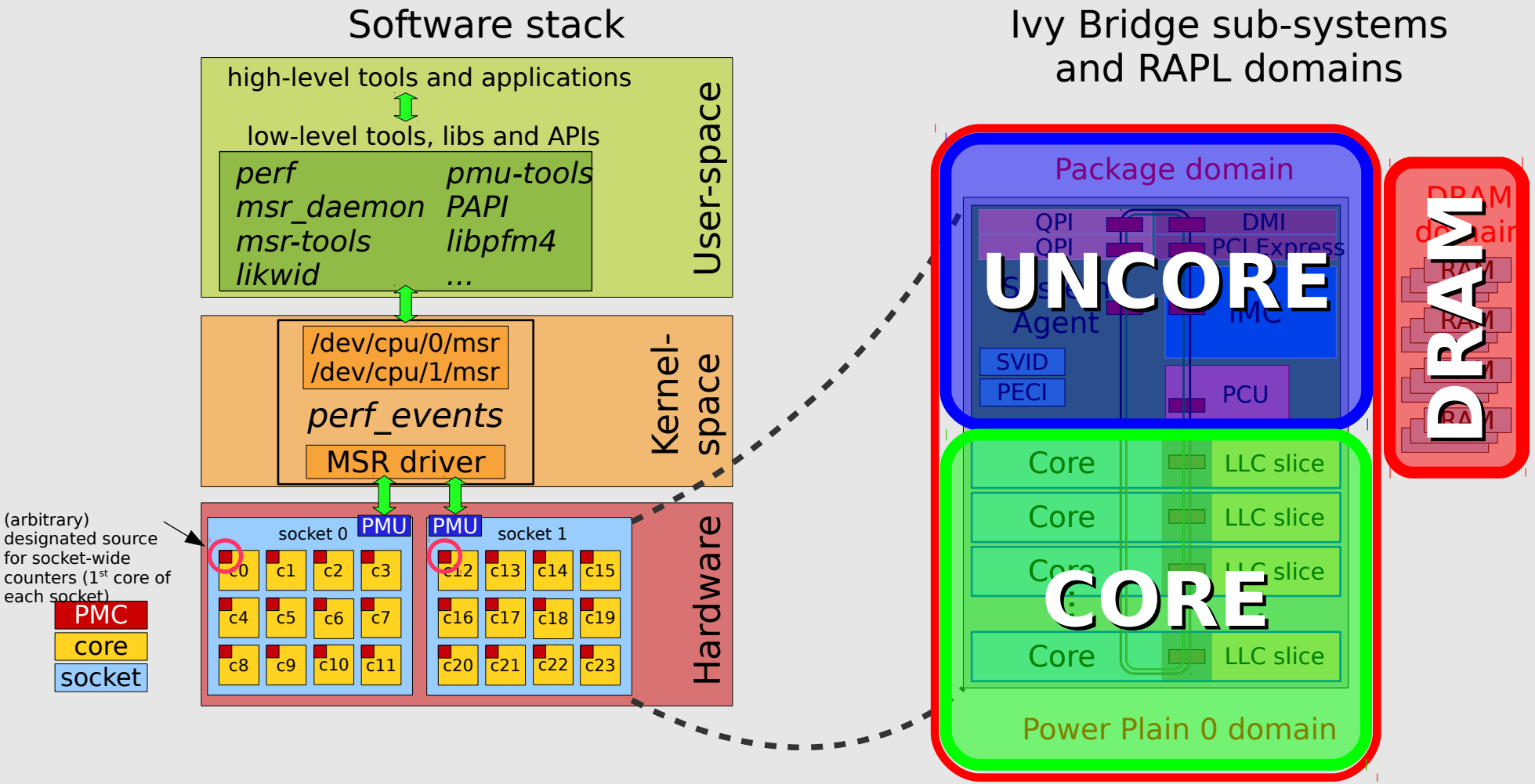
SISSA ICTP

Power Management and Acquisition Software

- *Eurora Monitoring Framework* (Micrel Lab's sw developed for Eurora)
  - **msr-statd**: MSR/RAPL acquisition software
    (C program revised, improved, now linked to *hwloc)*
  - **gpu-statd**: GPU status info
    (Python script interfaced to NVidia Management Library (*nvml*))
- linux **perf** in order to access and collect performance counters (used as well for the top-down characterization)
- **cpufreq-utils** in order to modify the CPU frequency scaling governor and CPU frequencies (require super-user's privileges)
- various ad-hoc parsers and wrapper scripts (bash, awk, sed, perl, python, C)

MHPC
Master in High Performance Computing

Software stack, CPU sub-systems & RAPL domains



SISSA ICTP

Testbed

C3E (Carnia Industrial park Cloud and Cluster Environments)

- 2 Aurora Chassis
  - 6 Aurora computing nodes, 24 cores each
  - 8 Aurora computing nodes, 24 cores & 2x NVidia K20 each

Each Aurora blade:

- Intel Xeon *Ivy Bridge* EP E5-2697 v2 @ 2.70 GHz, 12 cores, 30MB L3, in a dual-socket server configuration (*Romley*)
- 64GB RAM

Benchmarks to try..

- 2 well-known benchmarks
  - HPL (CPU-bound app.)
  - HPCG (memory-bound app.)

from scheduler perspective, the code is a black-box

MHPC
Master in High Performance Computing

Benchmarking and Analysis

- benchmarks tuning runs
- energy profiling of well-known benchmarks (HPL, HPCG)
- comparison through performance counters
- comparing performance and energy efficiency changing frequency and problem size (looking for a trade-off for memory-bound applications)
- comparing different BLAS libraries (Netlib, ATLAS, OpenBLAS, MKL) using HPL through all the aforementioned methods

SISSA ICTP

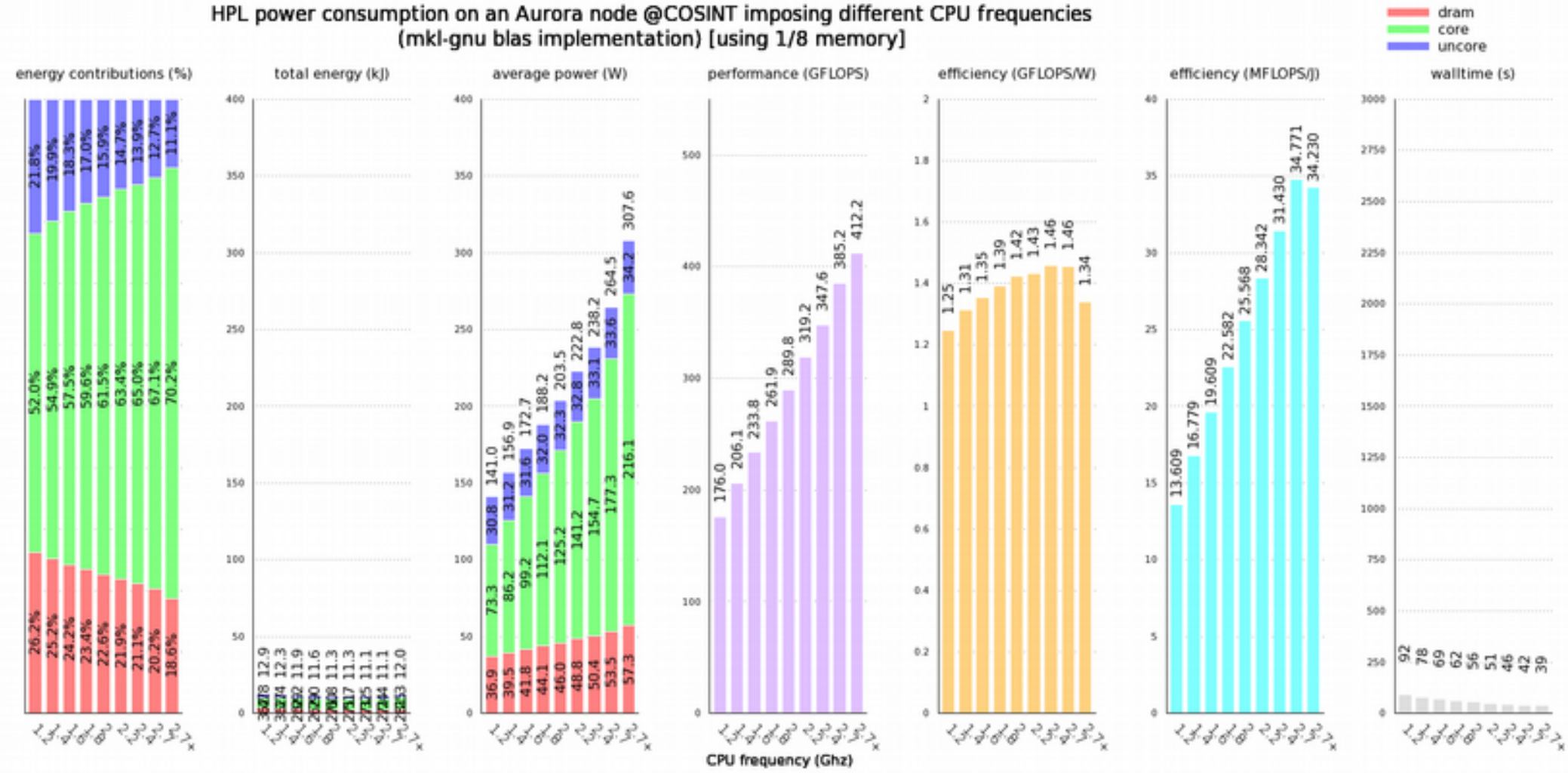
Why comparing BLAS libraries?

- several implementations available
  - NETLIB (reference library)
  - ATLAS (compile-time automagic optimization)
  - MKL (INTEL proprietary libraries)
  - OpenBLAS (free/open high-perf implementation)

- as the implementation changes, the achieved performance can be very different, energy efficiency must be different too (FLOPS/Watt - FLOPS/Joule)

MHPC
Master in High Performance Computing

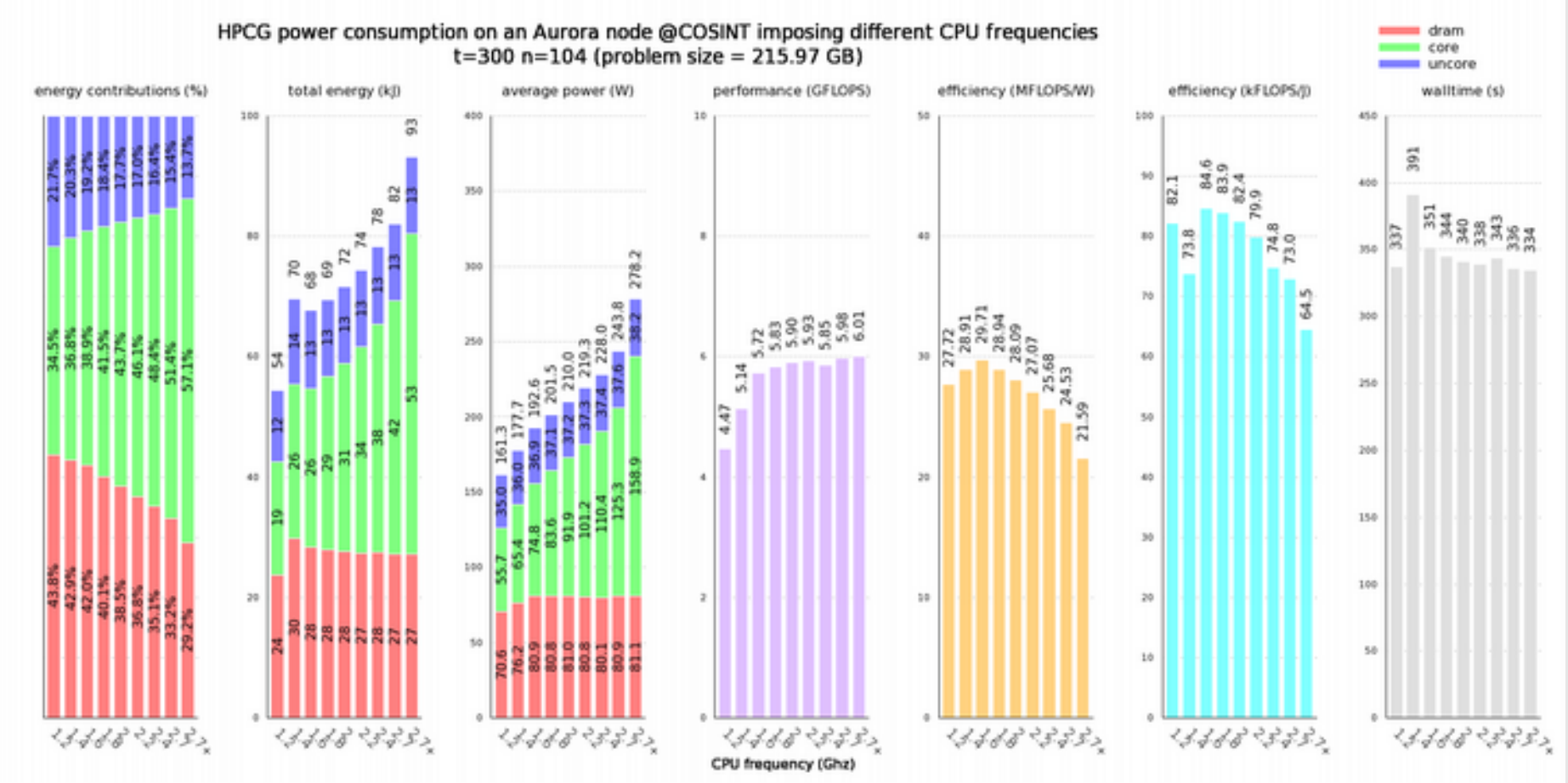
HPL+MKL: frequency scaling



SISSA ICTP

MHPC
Master in High Performance Computing

HPCG: frequency scaling



SISSA
ICTP

Observations

- **RAPL** and **performance counters** represent **powerful tools** for out-of-band/unattended profiling and monitoring (energy-aware scheduling)
- performance counters are **complex and difficult to handle** properly, several **hidden caveats** make them difficult to be widely exploited, further study is required

MHPC
Master in High Performance Computing

Final exercise for MHPC student

- Run HPL and HPCG on C3E and measure energy (see github account)

SISSA ICTP