Quantitative Methodologies for the Scientific Computing: An Introductory Sketch

- Alberto Ciampa, INFN-Pisa
  - alberto.ciampa@pi.infn.it
- Enrico Mazzoni, INFN-Pisa
  - enrico.mazzoni@pi.infn.it

Quantitative Methodologies for the Scientific Computing

- We want to evaluate: activity, efficiency, potentiality
- To account the costs: global costs, costs per Group/Experiment, costs per “produced unity”
- Power costs, investments, operative costs, maintenance, human (FTE) etc.: for this example we’ll consider Power costs

↓ To define:

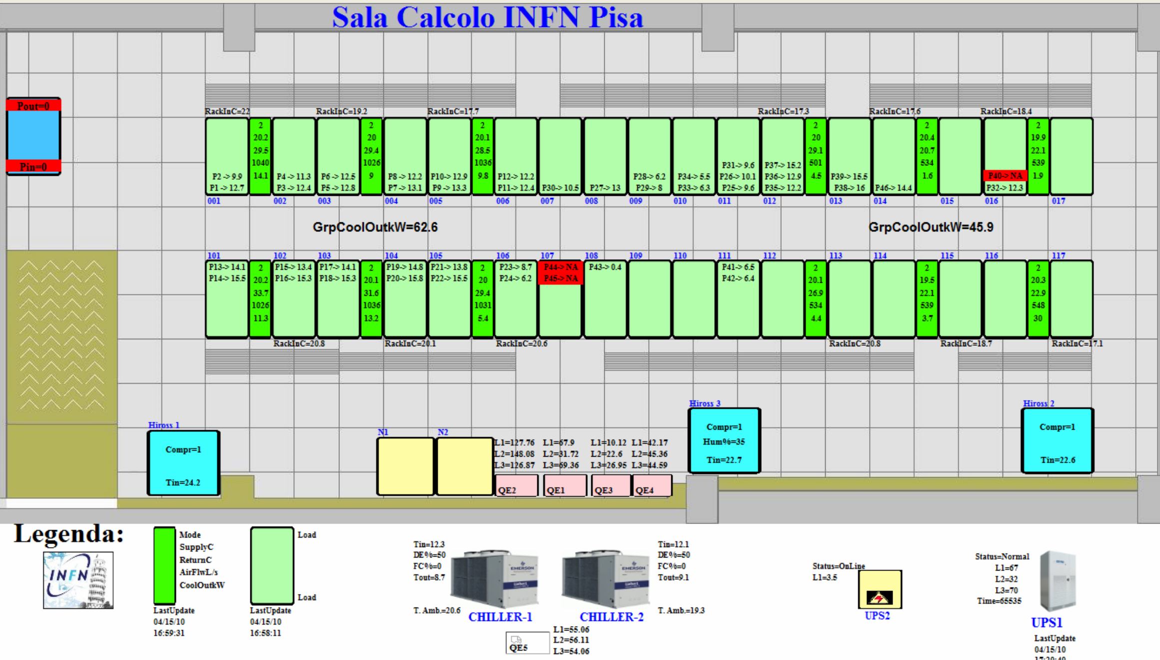
- Optimization strategies
- Forecast and scheduling
- New jobs and external work order evaluation

The INFN-Pisa Computing Center

Present situation:

- GRID ≈ 1900 production (WN) cores
  - CMS T2: 53%
  - Theophys: 39%
- National INFN Theophys Cluster: 1024 cores (under implementation)
- Services
- Storage ≈ 350 TB

The INFN-Pisa Computing Center



Plan

Roof

3D sketch

Introduction: A request from our Director

How can we account the various computing costs to the Groups/Experiments?

We started defining the “resources” and trying to define a model for their usage by the Groups/Experiments.

Introduction: The Resources and Production

Resources (examples):

- Rack space
- (production) CPU
- Network port
- Storage space
- Power
- Conditioning

A resource can be statically or dinamically allocated.

Production Model:

- GRID (and local queues): CPU dynamic allocation
- Farm dedicated to Group/Experiment: CPU static allocation

INFN Istituto Nazionale di Fisica Nucleare

Introduction: The Model

The final activity is the "production" measured (for instance) in day/core.

- The main resource is the computing core
- Some resources are "tailored": the power, air conditioning, network, rack space
- Some resources are naturally statically allocated: storage space

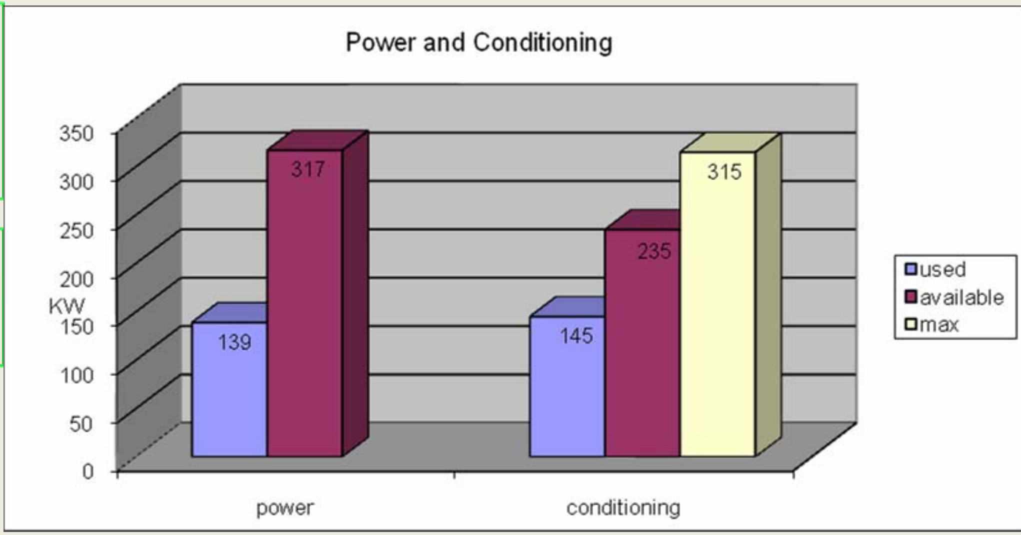
An Example: INFN-Pisa computing Center

RACK space: **34 rack**, 33 42U and 1 48U

Available power: 1380A+450A(roof) = 317.4KW + 103.5KW(roof)
Used power (mean): 602A + 228A (roof) = 138.5KW + 52.4KW (roof) = **43.6%, 50.6% (roof)**

Air Conditioning: 235 KW
Used (mean): 145 KW = **61.7%**
Max potential: 315 KW

LAN: 900 * 1GbE + 40 * 10 GbE
WAN: 1 Gb/s GRID + 400 Mb/s Sect.



May 2009 Data

The Model: Some Important Definition

- General Efficiency: % allocated (production) cores
  - GRID: % cores running a job
  - Farm: total of cores
- Specific Efficiency: % of UN cores
  - GRID: *cputime/walltime*
  - Farm: % (UN cores)/(total cores)

State of a core: SIWUN (System, Idle, Wait I/O, User, Nice)

Survey: Scientific Computing, 1/6/08-31/5/09

- **Computing Core**: **1567** = 1.332 (GRID) + 235 (Experiment Farm)
- **Computing power**: **2.35MSI2k (15.000 HepSPEC)**
- **Non Production Core for Scientific Computing**: **98** (GRID+dCache+GPFS)
- **Storage Space**: **300TB** gross
- **Max potential: 7.000 core + 1 PB storage** (quad core, 1 TB disk)

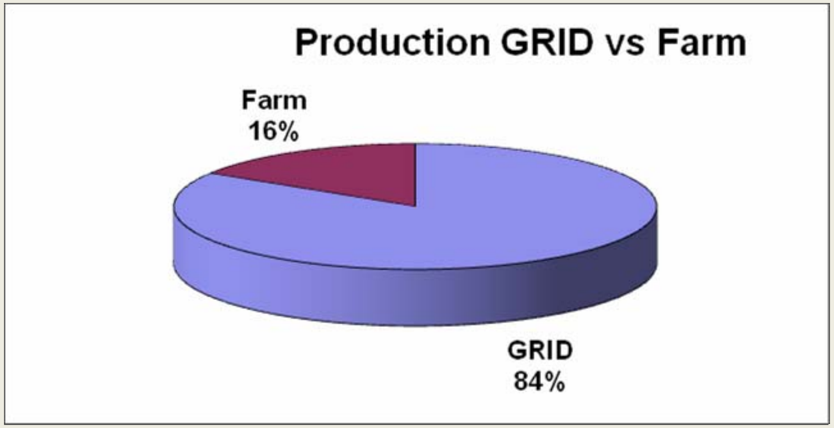
GRID

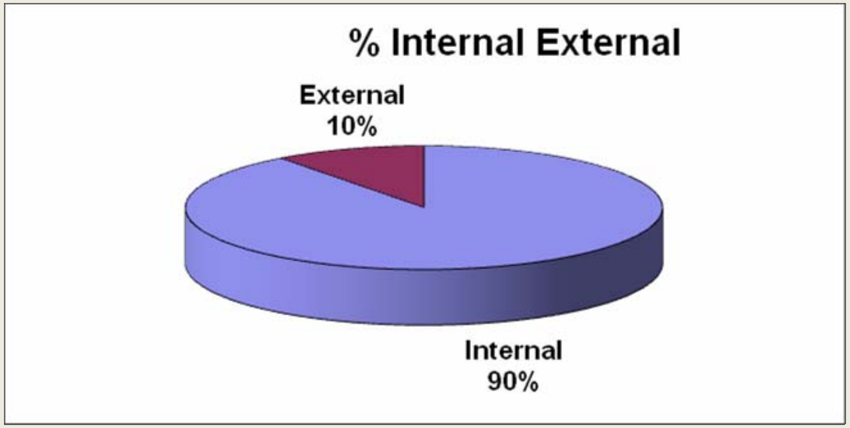
- GRID day-core: 350.971, that is 962 year-core.
- **Usage %**: **86%** (general efficiency), **75%** (specific eff.)=**65%** (total)
- Power usage (gross): 148.7 KW
- **Power consumption per day-core (gross)**: **3.44 KWh**

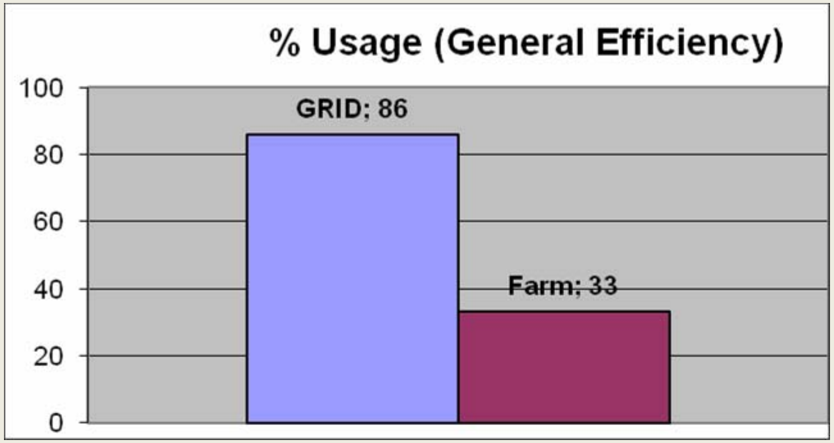
Experiment Farm

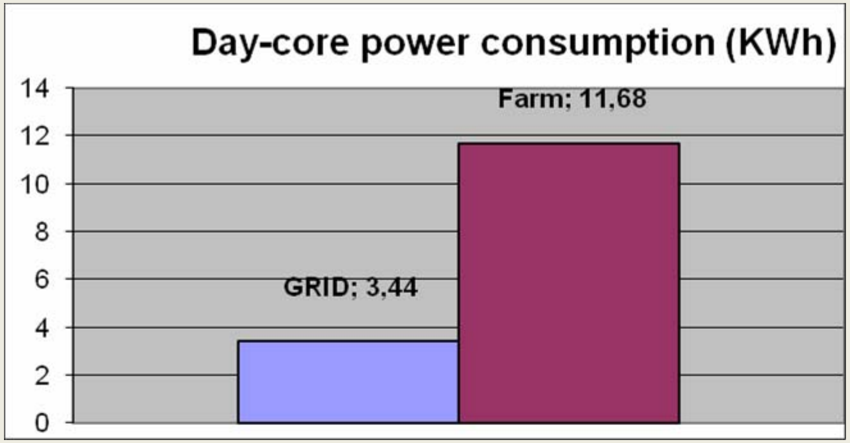
- Experiment Farm day-core: 66.759, that is 185 year core.
- **usage %: 33%** (general + specific efficiency)
- power usage (gross): 29.2 KW
- **power consumption per day-core (gross): 11.68 KWh**

GRID vs Farm

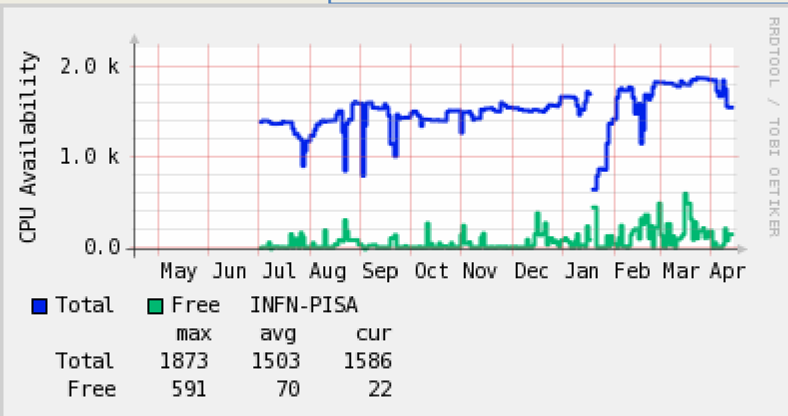

Production GRID vs Farm
Farm
16%
GRID
84%


% Internal External
External
10%
Internal
90%


% Usage (General Efficiency)
100
80
60
40
20
0
GRID; 86
Farm; 33


Day-core power consumption (KWh)
14
12
10
8
6
4
2
0
Farm; 11,68
GRID; 3,44

Example: GRID and local queues accounting

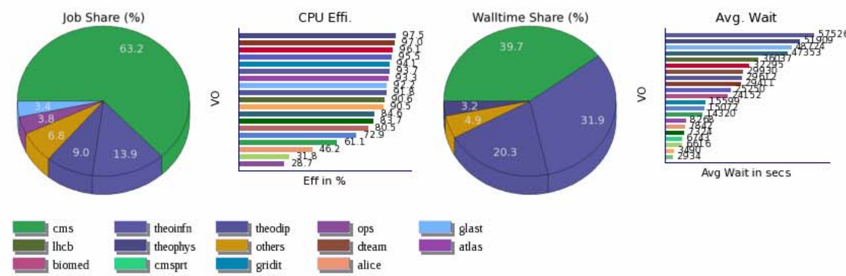


LSF Monitoring at Pisa

Current Status | Last 6 Hours | Last 12 Hours | Last Day | Last Week | Last Month | Last 3 Months | Last 6 Months | **Last Year** | Full Period

Jobs completed during the last year

VO/Group	Total Jobs	Succ Jobs	Succ Rate(%)	Walltime (sec)	CPU Time (sec)	CPU Eff(%)	Walltime Share(%)	Avg Wait (sec)
cms	1364055	1329695	97.48	18036835389	11024692448	61.12	39.72	14320
theoinfn	299574	286102	95.50	14487513843	13571751551	93.68	31.90	57526
theodip	193372	168052	86.91	9220591721	8463716179	91.79	20.31	29612
theophys	22676	21918	96.66	1455340159	1418289564	97.45	3.20	51909
glast	73492	72669	98.88	709181146	654073145	92.23	1.56	48724
lhcb	23706	23395	98.69	399296783	361665697	90.58	0.88	36037
biomed	15770	13840	87.76	255317715	62158397	24.35	0.56	24152
cmsprt	12895	12619	97.86	167513799	16802258	10.03	0.37	6743
compchem	3058	2935	95.98	126532879	122769647	97.03	0.28	29411
atlas	14874	14463	97.24	110002710	102680419	93.34	0.24	8268
cdf	1520	1506	99.08	90037999	76154596	84.58	0.20	47353
superb	2689	2658	98.85	86197670	77986699	90.47	0.19	3490
gridit	12580	11401	90.63	59160368	55661885	94.09	0.13	15599
theolong	3099	1466	47.31	52727207	50361823	95.51	0.12	25250
pamola	1721	1423	82.68	47033275	45214955	96.13	0.10	32295

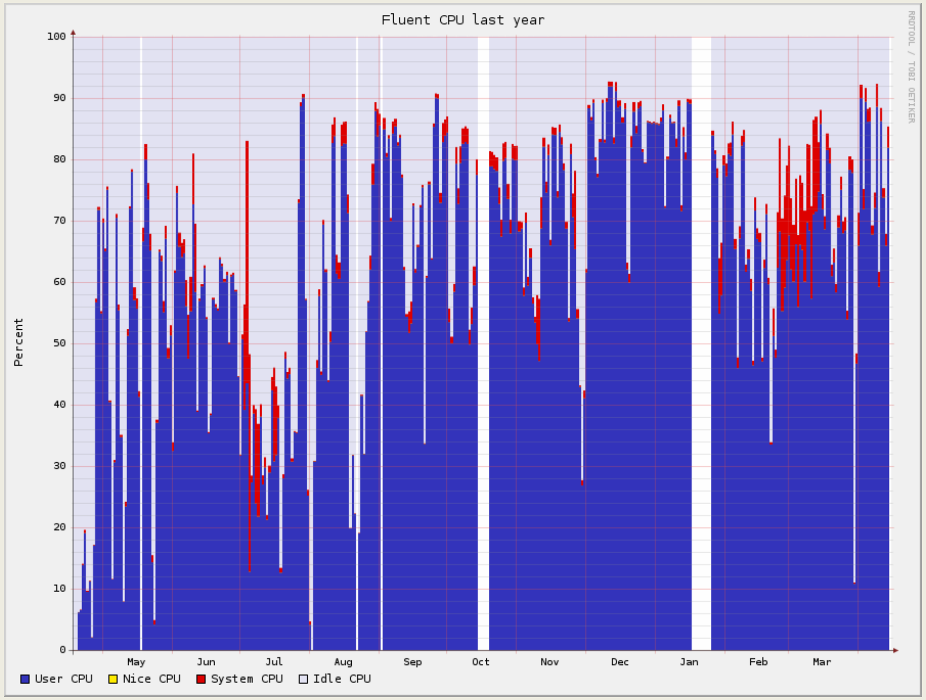


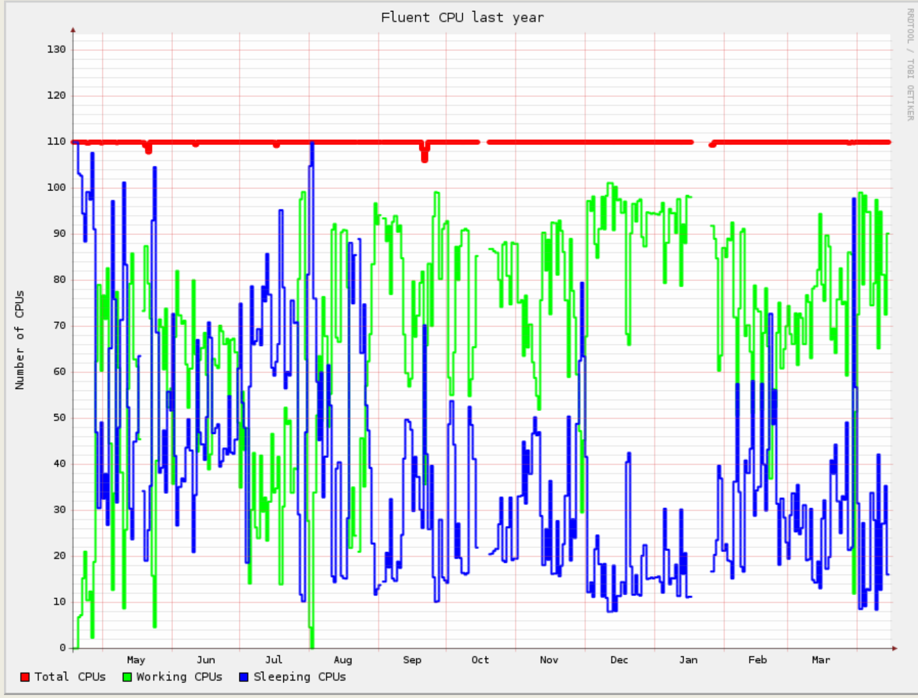
Last updated on 2010-04-15 18:31:08

Developed by: S. Taneja, S. Sarkar - INFN-

Example: Farm Accounting

Example of a farm dedicated to an external user

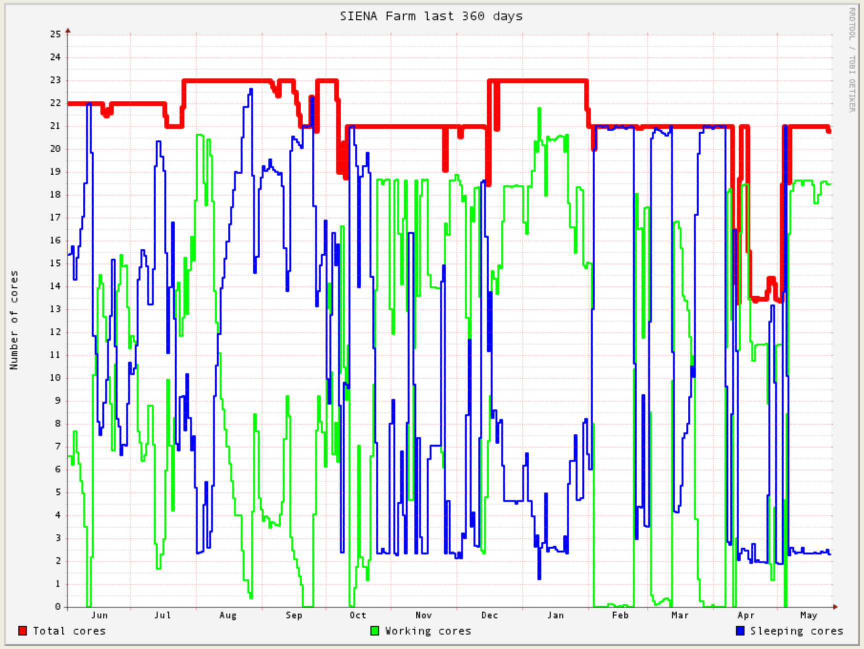


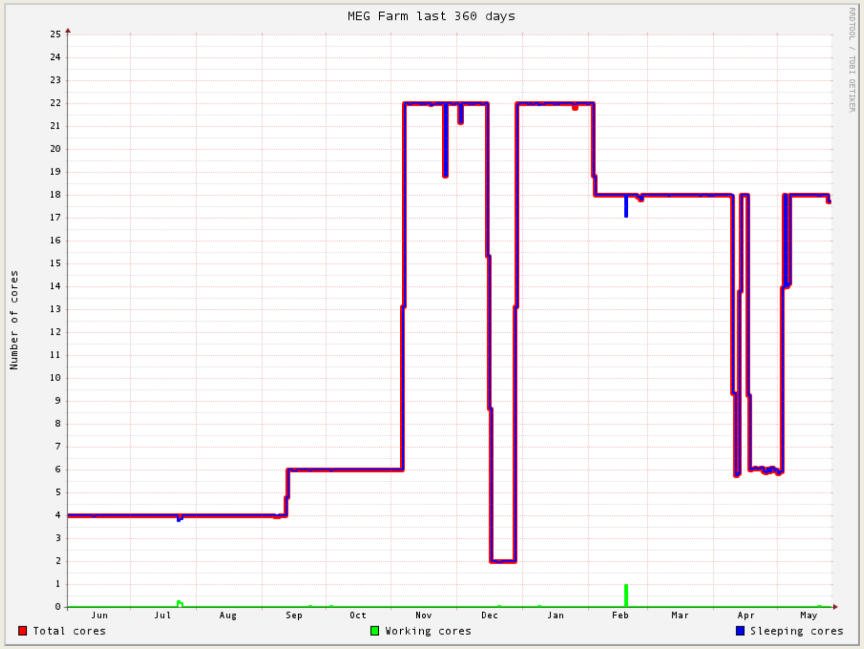


Raw data from Ganglia

Corrected data from Ganglia

Example of farm utilization





21 core
3.741 d/core, 49%

18 core
3 d/core, 0%

Data derived from Ganglia

Notes on the Methodology

- Scientific Computing as a Production Activity
  - It is not important, in general, the single production item, but the production flow
  - The quality of the system is measured as the production level (quantity) and the production efficiency:
    - *General Efficiency = % of infrastructure usage*
    - *Specific Efficiency = % of usage efficiency*
  - Excepted for specific cases we don't matter on what is produced, but how is produced: how much and how the systems are used
  - It is possible to, and we have to, measure the production costs, globally and for each "production line" (group or experiment): *Power*, investments, consumables, maintenance, FTE
  - It is possible, and we have to be ready to accept other job order, both internal (from our institute, mandatory) and external
  - **We do not pretend to cover everything** (interactive, T3, some farms)

Scientific Computing

- The Context
  - GRID (wn, middleware, SRM), Group/Experiment Farm and cluster
  - Storage for the above mentioned activities(+ disk server, SAN, Switch FC ecc.)
  - Data Center LAN (for SC), WAN dedicated to SC
- Usage Paradigms
  - GRID: only batch (LSF), resources shared among the VO, dynamic allocation following the requests (“fair-share”), accounting on usage (general efficiency). Including local queues.
  - Group/Experiment Farm/Cluster: the owner can freely access, reserved resources, accounting with static allocation non on usage base
- Users
  - GRID: all the VO accepted by INFN-GRID, “welcome” fair-share “, support (success job, efficiency)
  - Farm/Cluster: need good motivation, accounting including services (rack, network, high speed network, etc.), hosting

Quantitative Methodologies

- Resources
  - Global (%: SC and Service): space, power, conditioning, network, services*
    - * Shared Services (DNS, DHCP, Auth*, shared storage, etc.)
  - SC: Server, Storage
- Data: LSFMON (GRID), Ganglia (Farm)
- Measures:
  - LSFMON: (#core, #job, #queued) only mean, ∀ VO (#job, walltime, CPUtime) only Σ, #CPU
    - Integrals?
  - Ganglia: ∀ Farm (∀ Server (#core, %SIWUN)) with mean
    - Integral evaluation: not trivial
  - $CPUPower = (\text{RoomPower} + \text{Roof}) / \Sigma(CPUProd + CPUServ)$
    - Power vs Performance – CPU vs Core

Methodologies: GRID

Account core-time allocated to a specific VO (walltime), regardless exit status, no account if empty jobslot, account core-time SIW+UN (not considering specific efficiency)

- General Efficiency: $100(1-\int freecore / \int core) \approx 100(1-mean(freecore)/mean(core))$
- $GRIDPower = CPUPower \cdot \int CPUGRID \approx CPUPower \cdot mean(CPUGRID) \approx CPUPower \cdot mean(coreGRID) \cdot k$
- $Power/Day\text{-}Core = GRIDPower / \int (core - freecore) \approx GRIDPower/(mean(core) - mean(freecore))$
- Efficiency = (General Efficiency)·(Specific Efficiency*)
  - ** CPUtime/Walltime, both global and per VO*

Methodologies: Farm

Total cost independent from usage level (no specific efficiency= resources always allocated)

- General Efficiency: 100(∫*core*UN / ∫*core*)
- *FarmPower = CPUPower·∫CPUFarm*
- *Power/Day-Core = FarmPower/∫CoreFarmUN*
- Efficiency = (*day-core working*)/(*day-core allocated*) = ∫(*#core·%UN*)/∫*#core* *
  - * both global and per Farm

Further Info

- http://web.infn.it/CCR/index.php/note-interne-ccr/69-note-interne-ccr-2009/271-ccr-332009p-calcolo-scientifico-prime-metodologie-quantitative-per-un-ambiente-di-produzione sorry, in italian, but for any question you can contact:
  - Me (alberto.ciampa@pi.infn.it)
  - Enrico Mazzoni (enrico.mazzoni@pi.infn.it)