PIC
port d'informació
científica

Systems Performance and Cost Modeling Working Group

Overview, Mandate, Activities and Status

J. Flix (PIC/CIEMAT), M. Schulz (CERN), A Sciabà (CERN)

Joint WLCG & HSF Workshop - Naples, Italy - 28th Mar. 2018

WLCG Worldwide LHC Computing Grid & HSF HEP Software Foundation

Motivation I

Considerable gap between our best estimates of the required and the available future computing resources (HL-LHC)

- HL-LHC operating parameters increase computing scale on many levels
- Flat budgets (likely) to stay → rely on technology advancements (~15%/year for CPU and ~25%/year for disk)

To overcome this problem, **paradigm changes will be needed**

- Integrate “cheaper” and opportunistic resources (Clouds, HPCs, ...)
- Make use of more specialized hardware (GPUs, FPGA, etc.)
- Find more efficient algorithms (tracking, simulation, analysis, ...)
  - *Find optimal trade-off between precision, computing time and data size*
- Deploying a big data cost-effective pool/federation to host the bulk of the data (“Data Lake”)
- Optimize data retention policies
  - *Keep only 1 or even zero copies of RAW data?*
- Make sure that the computing resources are an optimal match for the loads and can be fully utilized
  - *Reduce inefficiencies to a minimum*
- ...

Motivation II

We need metrics that allow us to **characterise the resource usages of workloads** in sufficient detail so that the impact of changes in the infrastructure or the workload implementations can be quantified with a precision high enough to guide design decisions towards improved efficiencies

- *Resource utilisation* of the workloads in terms of fundamental capabilities that computing systems provide, such as storage, memory, network, computational operations, latency, bandwidths…
- *Performance assessment*
- *cost efficiency*: an approach for sites to map computing requirements to local costs (highly desirable)
    - *This can't be achieved at a global level, since the conditions at different sites are too different (in-kind contributions!), but a model should be constructed in such a way that this mapping on a local level can be done fairly easily*
- Decisions on the evolution of workloads, workflows and infrastructures might impact the quantity and quality of *human resources* required to build and operate the system

It is important that a **cost and performance model** at the system level takes these adequately into account to allow to optimize the global infrastructure cost into a constrained budget

Study in Constant Evolution

There are **many unknowns** whose effect we must be able to estimate

- Understanding of the impact of detector upgrades and HL-LHC conditions is continuously improving
- Software stacks continuous evolutions (processing softwares, DM tools, ‘middlewares’, …)
- Technology improvements (roadmaps and new technological paradigms)
- In our distributed model, we do profit from many unknowns in-kind costs…

Many stakeholders should be involved to monitor/study all of these effects in detail: from experiments to site experts

Systems Performance and Cost Modeling Working Group

Kick-off meeting in 9th November 2017 [**agenda**]

- Discussed mandate, overall roadmap and organization
- First discussion on how the work could be best organized (sub-working groups)

WLCB MB endorsed the creation of the group by 14th November 2017 [**mandate**]

- **Conveners:** J. Flix, M. Schulz, and A. Sciabà
- **Members:** 35 active members → **wlcg-SystemPerformanceModeling@cern.ch**

Active participation from workload, workflow and framework developers, people who plan, engineer and operate IT systems and people who put all this in global WLCG

- *Cross-links:* given the focus on computing for (HL-)LHC and an established community driven governance, this WG is best established as a "WLCG Working Group". However:
  - *The scope is not limited to WLCG and is of potential interest for other experiments relying on a largely distributed computing infrastructure, like (but not limited to) Belle II or SKA*
  - *reporting links between WLCG, HSF and HEPiX will be established*

Mandate

WLCG, HSF, HEPIX and the Systems Modeling Working Group (reviewed by MB Nov 14 2017)

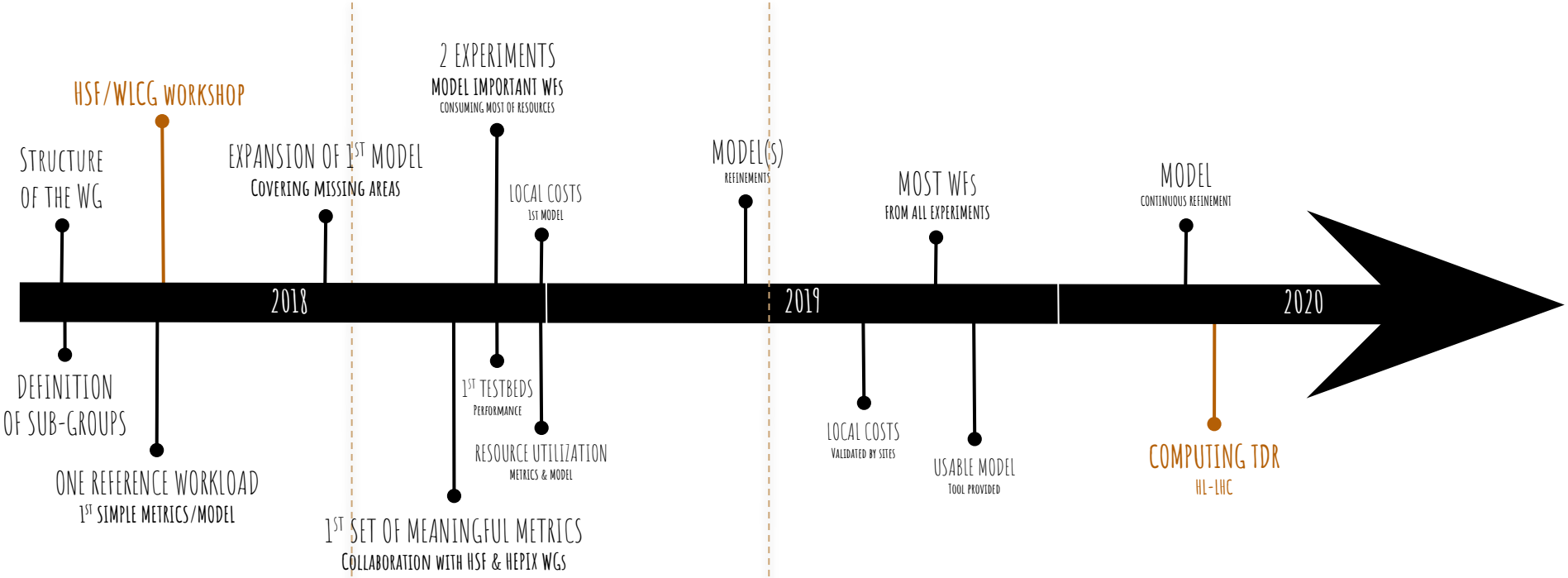
This working group depends on active participation from workload, workflow and framework developers, people who plan, engineer and operate IT systems and people who federate all this into a global infrastructure. To ensure that this activity is reflecting the understanding of these vital groups of experts it is helpful that beyond the informal cross links created by members of the working group, reporting links with all three entities will be established. With WLCG's history of tracking a large number of working groups, the focus on computing for (HL-)LHC and an established community driven governance, this working group is best established as a "WLCG Working Group". The scope is not limited to WLCG and is of potential interest for other experiments relying on a largely distributed computing infrastructure, like (but not limited to) Belle II or SKA.

Mandate (reviewed by MB Nov 14 2017)

- Bring together workload and infrastructure experts from sites and experiments to agree on common suitable metrics to describe the interrelationship between workload resource needs and infrastructure characteristics.
- Identify and agree on a set of reference workloads and meter them in different environments as input data for a model.
- Build models and verify them by predicting resource usage of the reference workloads with respect to changes in the execution environment.
- Develop a strategy/methodology for mapping the model predictions to local costs and verify this approach by applying it to a small number of different sites.

These steps are not necessarily sequential and an iterative approach is needed to end up with usable metrics and models that can over time follow the changes in our environment.

Roadmap [preliminary]



SHORT | MEDIUM | LONG

(main) WG tasks (by Mar. 2018)

[#1] Revision of most important workloads for each experiment
[#2] Packaged versions of the most important workloads
[#3] Definition of properties best characterise a workload
[#4] Draft a cost evaluation process
[#5] Compile a list of relevant performance analysis tools
[#6] Set up a distributed testbed to run tests
[#7] Start a simple resource calculation model

[#1] Revision of most important workloads for each exp.

Most important Workloads → Those that consume the **largest fraction** of the resources used by an experiment (current and future)

Information available from the experiments and **regularly reported** to the WG

All of the 4 experiments have already described:

- Current status and 1st expectations for Run3 and HL-LHC
- Workloads characteristics: **CPU intensive** | **CPU+I/O** | **I/O intensive**

Ex. MC Simulation
GEANT4

Ex. Digitization
Ex. Reconstruction
Ex. Stripping

Ex. SKIMMING
Ex. MERGING

- Provided task configuration(s), performance metrics tools, setups to run isolated tasks, …

  → ***Group members have already verified the 'recipes'***

→ David talk

→ Johannes talk

[#2] Packaged versions of the most important workloads

In collaboration with the **HEPIX CPU Benchmarking WG**

Several **containers** available (atm, for ATLAS and CMS), for different Workloads, experiment software versions, and OS

The experiment ‘recipes’ to run Workloads are regularly **packaged**

- In particular, those that are described in WG task **#1**
- CVMFS mounts are needed (for experiment software)
- Typically, data is read with no credentials (local files, or remote reads via XRootD)
- Intermediate files created by the user(s) running the image(s)

This allows to run the most important experiment workloads in a **simple way**

- This might be integrated into HammerCloud to run at scale (if needed)

It needs **revisions** (using more recent experiment softwares, for example)

[#3] Definition of properties best characterise a workload

Those that allow to calculate how many resources are needed and of what type

- *that sites can use for their purchasing plans / that can be used as input for a model*

Metrics can be useful for software developers who look to improve the resource usage of their software

- How to measure the metrics? How to identify the most critical ones?
- Metric external availability and/or calculation from the application?
- Is the metric available at user level or requires root access?
- Does it requires code instrumentation?
- Its collection has a performance impact on the application?
  → It requires lot of discussions/prototypes to get the set of workable metrics

Areas: CPU, Memory, IO, Network, Application Throughput

The plan is to gather all of these metrics from the workload job examples

- Prmon tool and other tools

→ Gareth talk

[#4] Draft a cost evaluation process

Goal: Outlining a method to calculate the cost of hardware to expand the site capacity to hold the most important workloads, where the cost is split into several components

1st simple exercise: given the resource needs of a workload, close to an existing one, a few sites are describing how they would estimate the cost of running 10^6 realizations, based on their current infrastructure

- *Understanding the main workload characteristics used for these estimates*
- *Identifying what is missing*
- *Discuss the approaches to reach an estimate (identification of commonalities)*

Note: Nordic Tier-1 review and costs evaluation report (2015)
→ https://wiki.neic.no/w/ext/img_auth.php/6/66/NT-1-Evaluation-report.pdf

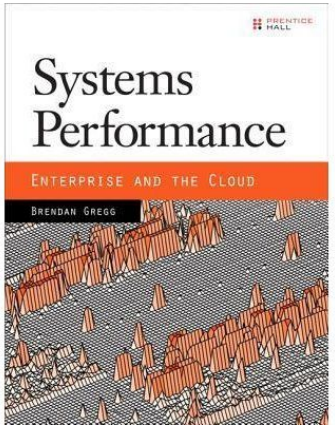
→ Catherine talk

→ Renaud talk

[#5] Compile a list of relevant performance analysis tools

Look at commonly used performance analysis **tools**

→ figure out which ones are the most useful to characterize the workloads



Focus on:

- Common Linux system commands
- Tools assessment (*perf*, *bcc*, *SystemTap*, *graphical tool for memory accesses*, ...)
- Automated Metric Collection (*node_exporter*, *Prometheus*)
  - *Formats & publishing*
  - *Configuration*

On the longer run, this will be expanded to tools and **practices**

- build a knowledge base that our community can use and that establishes common practices
- Compact, but useful for us, set of wrappers that allow less experienced members to study and evaluate the systems and codes [prmon, trident,...]

→ Valassi talk

[#6] Set up a distributed testbed to run tests

A collection of machines, possibly at different sites, should be made available to run workloads in a **controlled environment**

It should be easy for WG members to get access

→ root access might be needed for some performance analysis tools

Ongoing discussions from some site members on how to provision the **testbeds**

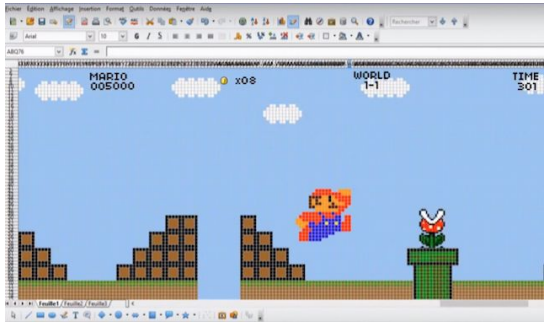
→ **Several sites** already offered support for it
→ **CERN testbed is being used**, from day one (open to anybody requiring access, indeed)

[#7] Start a simple resource calculation model

Taking inspiration from the **complex spreadsheets** that experiments use to calculate their resource needs

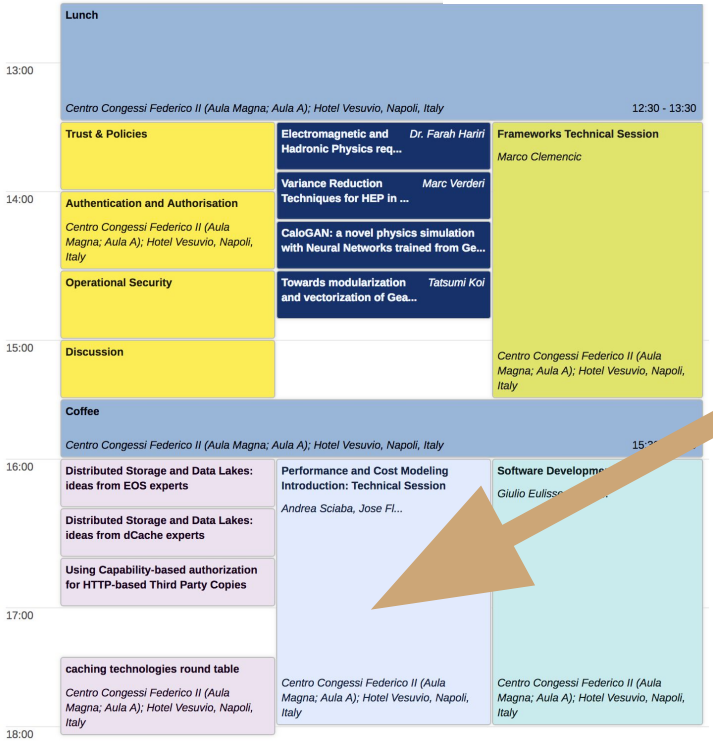
Create a simplistic model that can be progressively refined

Not much progress yet… but it is agreed that it is much better that this is provided by means of a **code** (K. Bloom example), rather than continuing with the spreadsheets



→ Sartirana talk

Performance and Cost Model @WLCG/HSF Workshop



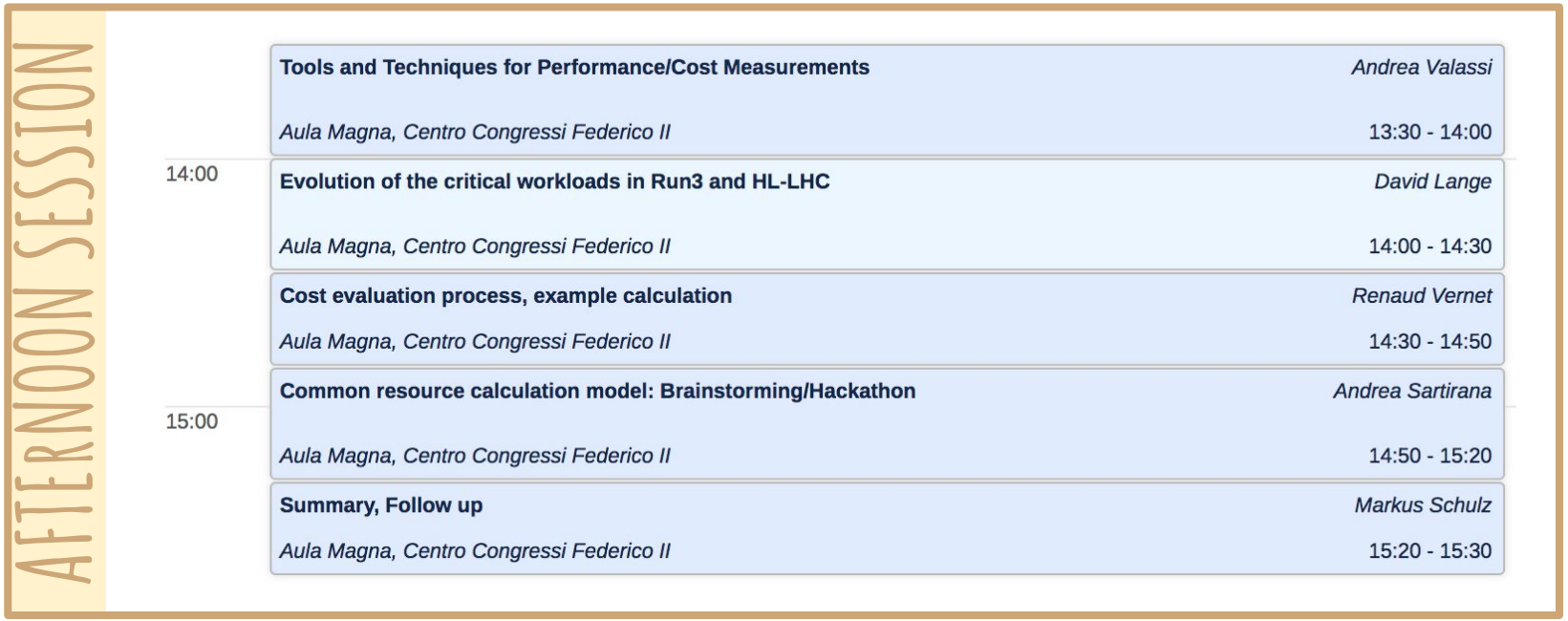
90' IN THE MORNING
(GENERAL AUDIENCE)

120' IN THE AFTERNOON
('WORKING' SESSION)

Performance and Cost Model @WLCG/HSF Workshop

MORNING SESSION

Performance and Cost Model @WLCG/HSF Workshop



Notes

We do need **volunteers** to take notes for today's Cost Model morning and afternoon sessions

At least 2 persons for each session

Gdoc link to include the comments lively:

https://docs.google.com/document/d/1Sc6npKa5l6RK0ifDee3MedjGRICesPrq5oXHBD8xf6U/edit?usp=sharing

[Names to appear here]

Conclusions

Resources in WLCG will be more and more constrained in the next years

- Attention is focusing on increasing efficiency and cost-effective scenarios

Performance studies aimed at a more detailed characterization of workflows are needed

A model to accurately estimate resource needs and best allocate spending is needed

- Even if the model is not very precise, a common approach would be extremely valuable

WLCG has formed a working group to organize this long term activity

- Close collaboration with HEPiX and the HSF is essential and planned
- Good (initial) progress so far - regular meetings scheduled

*** Want to volunteer? contact the authors of this talk!**


LOT of WORK AHEAD

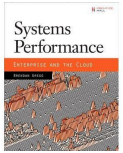
Useful links / References

Twiki: https://twiki.cern.ch/twiki/bin/view/LCG/WLCGSystemsPerformanceModeling

WG meetings: https://indico.cern.ch/category/9733/

Mailing list: wlcg-SystemPerformanceModeling@cern.ch

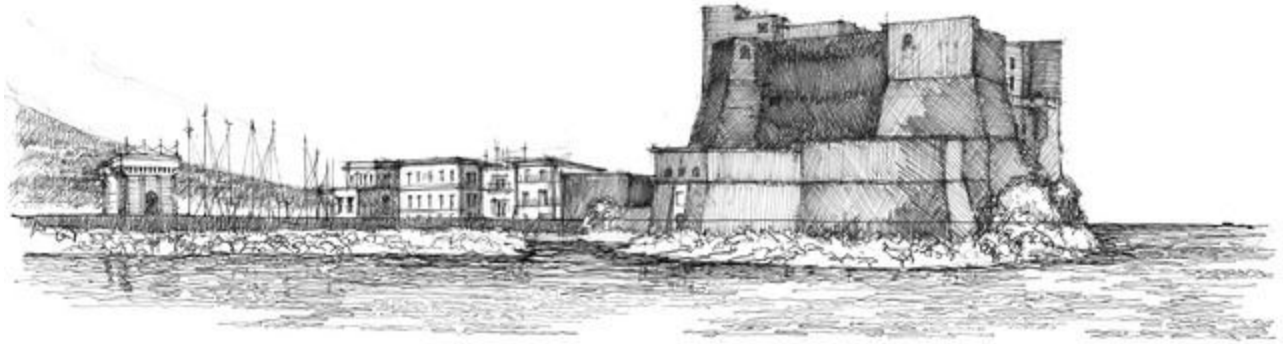
HEPiX Fall 2017 WG talk: http://cern.ch/go/7Z6h



Essential book: Brendan Gregg, Systems Performance, Prentice Hall

Brendan Gregg's blog: http://www.brendangregg.com/blog/index.html

B. Panzer's tech trends: https://twiki.cern.ch/twiki/bin/view/Main/TechMarketPerf



Questions?