WLCG-LHCC meeting - CERN - 9 May 2017

CMS

Liz Sexton-Kennedy (Fermilab)
D. Bonacorsi (Univ. of Bologna / INFN)

(on behalf of CMS Sw/Comp)

Outline

Today (as from indications received):

- usual activity overview (since last LHCC meeting) → **omitted this time**
- outcome of C-RSG (Apr'17) for CMS
- CMS towards the HL-LHC Computing TDR
- how to track resources (and plan accordingly)
- analytics

Timeline of discussions on 2017/18 resources

CRSG Spring'16: first resource requests for 2017

After an extraordinary 2016 for LHC, CMS prepared and submitted revised requests for 2017/18

- 2017: after massive mitigation work, down from +40% to +20% (on average across resource types)

CRSG Fall'16: CRSG **scrutinised and fully endorsed the 2017 requests as they were** (2018: not scrutinised)

RRB Oct'16: recommended the CRSG numbers for 2017. Dialogue on 2017 with FAs started.

CMS continued to run ops through Winter 2016/17

LHCC document submitted on Feb 6th. Last LHCC meeting was Feb 21st.

Situation as of April 1st, 2017

	ALICE	ATLAS	CMS	LHCb
CERN CPU	0%	0%	0%	0%
CERN disk	0%	0%	0%	0%
CERN tape	0%	0%	0%	0%
T1 CPU	-8%	-12%	-14%	-4%
T1 disk	-14%	1%	-21%	-6%
T1 tape	-1%	-7%	-24%	-3%
T2 CPU	-24%	-13%	-7%	27%
T2 disk	-28%	-7%	-22%	-30%

(more later)

Source: https://wlcg-rebus.cern.ch/apps/pledges/summary/

NEW since last LHCC

CRSG Spring'17: CRSG **scrutinised and fully endorsed the totals of the 2018 requests**

RRB Apr'17: Dialogue on 2018 with FAs started.

A reminder

We have been working in 2016 to explore most (if not all) reasonable ways to reduce resource needs in 2017/18

- just one slide of summary in the following

Main drivers

already shown at last LHCC

in the updated resource requests model for 2017/18

Reconciliation of 2016 into the model, given the final LHC delivered data in 2016

- we did not get the Msec we were supposed to get as from the July-updated projections

Further step towards **using less the RECO data tier**

- now used basically only for prompt data detector commissioning
  - we still had fraction of RECO on tapes, we stopped it in 2016 and we survived

Reduction of processing times for data and simulation processing

- most prominent is the adoption of the "pre-mixing" paradigm in production
  - simulation processing time driven down by ~50%. Drawback is premixed samples lib needed, but doable

Reduction of number of AOD replicas + more aggressive MiniAOD adoption

Move away from RECO and (partially) AOD data tiers for **Phase-II processing**

- it has become more space consuming → need to mitigate: go to mostly MiniAOD after an initial period

Aggressive cleaning of (mostly Run-I) data on **Tier-0/1 tapes**

- Summer/Fall 2016: see next slide

More aggressive GEN-SIM custodial policy

- processing chains; <100% GS saved on tape; smaller average size

A reminder

We have been working in 2016 to explore most (if not all) reasonable ways to reduce resource needs in 2017/18

- just one slide of summary in the following

IMPORTANT: once a change was studied and identified as meaningful and yielding a reasonable benefit at affordable costs, it was implemented in the model already in 2016, it already mitigated massively the 2017 resource requests, and the same benefits apply since then also to the 2018 requests

The work continued in 2017:

- LHCC document delivered in early 2017 (as reported at last LHCC meeting)
- also in the context of initiatives launched by the CMS management
  - ECoM-17 working group (Evolution of the Computing Model), expected to deliver through Summer'17

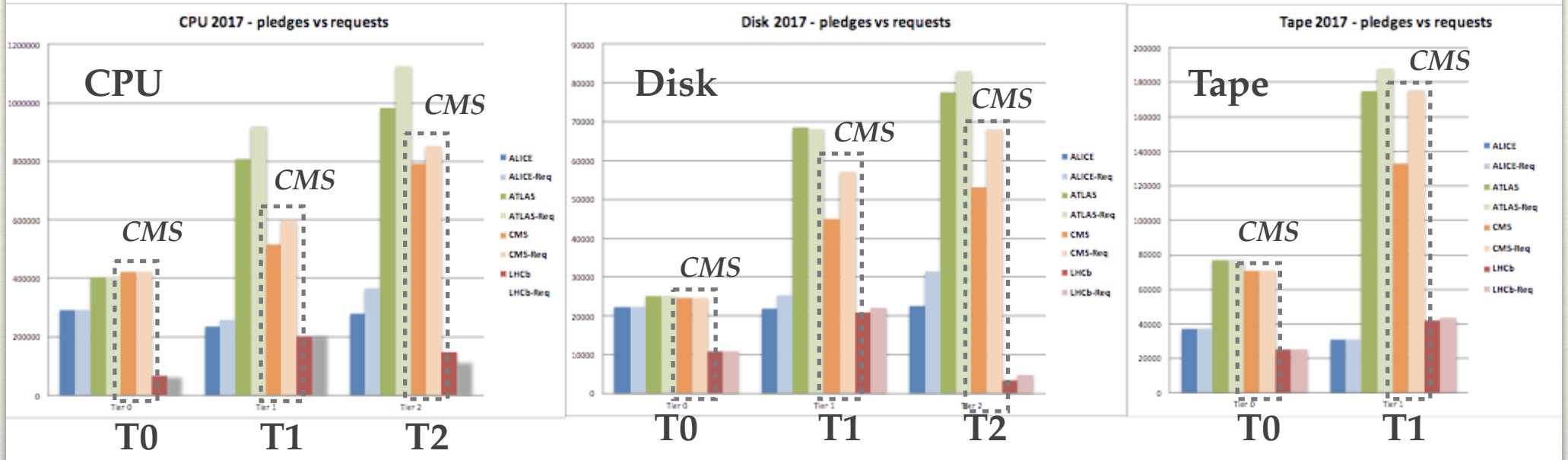
The LHCC doc in one table

already shown at last LHCC

Full doc sent by mail to LHCC referees on Feb 6th

Areas definitions (by LHCC)	Priority	Status	Few remarks
Optimization of workflows	very high	DONE	**Premixing** mode for high PU simulations at large scale. Deployment Nov 2016, used in production for MC DigiReco for Moriond'17. Local IO reduction + reduce 2x amount of CPU needed for DigiReco → minimise the CMS requests for CPU in 2018. Performed studies to exclude statistical biases. Pressure on remote access ops/monitoring (but a robust data federation is beneficial anyway).
Technology improvements	high	R&D	Potential to reduce complexity by large factors, but in which technology area and the timeline are largely unpredictable. **GPU** integration, **HPC** centres exploitation, opportunistic **cloud** extension of WLCG centres, overall orchestration of diverse resources, new FA mechanisms to offer these as pledges are all aspects to consider. Integration efforts (and manpower) not negligible.
Data / CPU / Tape management	very high	partially DONE	CMS computing model evolved towards higher flexibility in LS1. Main workflows can be submitted (almost) at any Tier level. Commissioned **processing chains** for better streamlining of global processing efforts (e.g. GEN to miniAOD can be run as a single step) → save CPU and especially tape. A limitation comes from being impractical for all GS, but very useful for a fraction of them.
Triggers thresholds tuning	medium	not DONE	Negative impact of rate reductions on physics output is potentially large. May be justified resource-wise only if reductions are sizeable. Rough estimates indicate that 1kHz → 800 Hz yields relatively modest savings, and put some physics programs at risk. Requires careful scrutiny and guidance by ECoM. In general, CMS would not suggest to pursue this path.
Amount of simulation	medium	not DONE	**MC/data ratio** tuned at 1.3 in the CMS resources model. Recently needed to do more than expected (both 2015 and 2016). Rough calculations for a 130%→100% reduction in 2018 yield savings of -8% (CPU), -3% (disk), -3% (tape). Extreme caution needed to avoid impact on physics by such reduction. New assessment for optimal tuning is in the ECoM mandate.
Parking / Delayed processing	medium	1. DONE 2. not DONE	Distinction between **delayed processing** and **parking+scouting**. CMS can do (and does) the latter, but would discourage to pursue the former. Gain of parking 200 Hz to be rereco'ed later would be quantitatively similar as the estimates for Triggers thresholds tuning above, with no gain on tape space as RAW will still be written. Devastating impact in the former case for B physics, for instance
Copies / Formats versatility / Analysis Frameworks	high	almost DONE	**MiniAOD** format introduction, ~8x gain in size, used by ~80% of the analyses today. Larger adoption is planned, but it will take time. → reduce the disk needs. Caveat: during the transition, miniAOD plus a fraction of AODs need to stay on disk to support all analyses. Dynamic use of storage space, load on Ops, mitigated by more automation. Remaining need for AOD must be small.

2017 Pledge situation



Not all is deployed yet for 2017 – a few delays
Full resources expected by June

CMS
CMS-Req

Baseline for CMS is to be able to do our computing operations on a pool of (pledged) resources that is the **dark orange** in the plots above

One important remark

CMS fully acknowledges the **help from FAs** that - despite financial constraints - understand that we mitigated the request needs as much as reasonably possible while protecting the CMS physics program, and are finding ways to support CMS towards (and beyond, in some cases) the pledges

This will be crucial for CMS operations in 2017, and very important for 2018 too

- far from obvious to plan on amount of resources that are beyond pledges: working (in WLCG and in CMS) to find optimal ways to track their availability

Will 2017 be a problem?

We should be able to handle the current year

- with pledges + additional efforts from few FAs + flexibility in the model
- the challenge was in 2016. We acted proactively in Summer 2016 and attacked it. **All mitigations we explored were aggressive, and have been implemented in a stable manner in the model**

already shown at last LHCC

More details per resource category:

- TAPE → heavy cleaning in 2016; "do it once and do it big"; pledges fully OK for the T0 (and apparently no problems there).
- CPU → the least critical resource type; -7/14% at Tier-2/1 wrt pledges; flexibility of the model helps here, as well as efforts similar to HEPCloud; should be able to use over-pledge resources at T2s in some regions; a coherent resource utilisation planning across diverse sets of computing environments (opportunistic, HPC) will be a hot topic in 2017
- DISK → drastic actions (# replicas, refocus main data tier from AOD to MiniAOD) will be the key. Caution while being in the transition, which (by experience) means using some more disk for a while before using less..

The CRSG exercise to scrutinise 2018

Pledged Resources in the recent years

4

Study of pledged resources

- Plot pledges from 2013 up to 2017 as they are in REBUS as function of the year;
- Fit each plot to measure the actual average increase year per year;
- Display 2018 requests, not used in the fit, and compare them to the fitted value extrapolated to 2018.

Reminder, flat budget assumptions: 20% increase for CPU, 15% for disk and tape space at constant budget

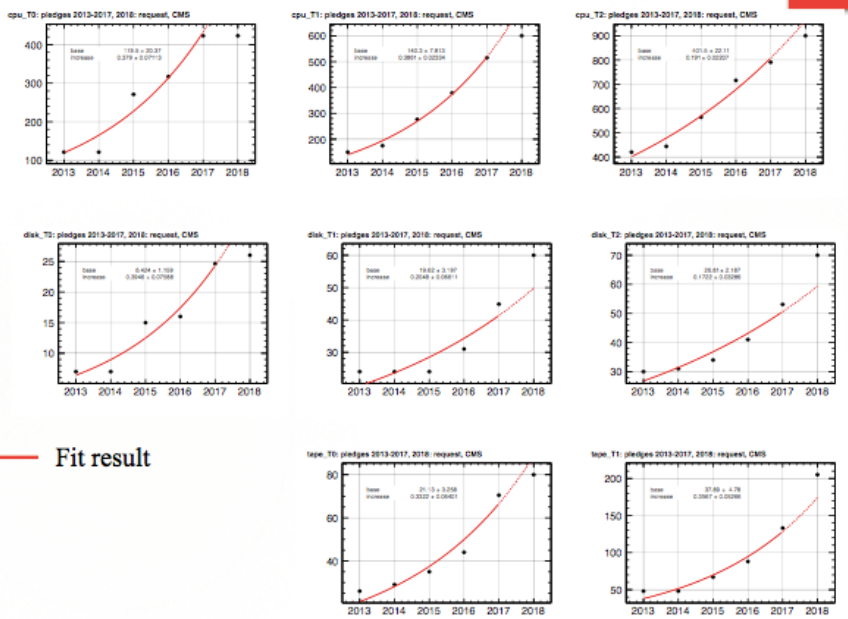
D. Lucchesi for C-RSG CERN-RRB-2017-057

April 25, 201

This is the outcome for CMS (a couple of plots in next slide for better readability)

CMS History

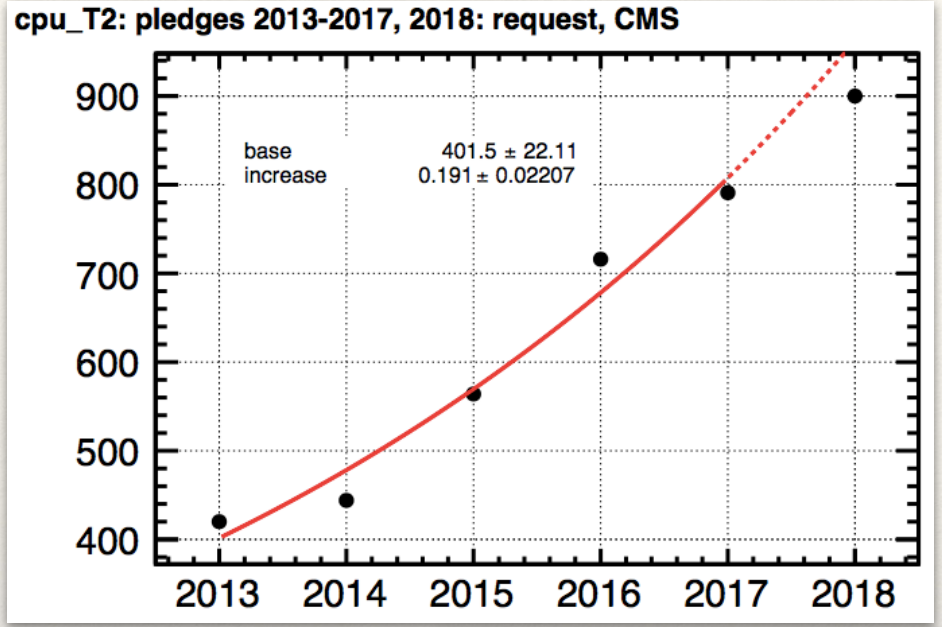
7

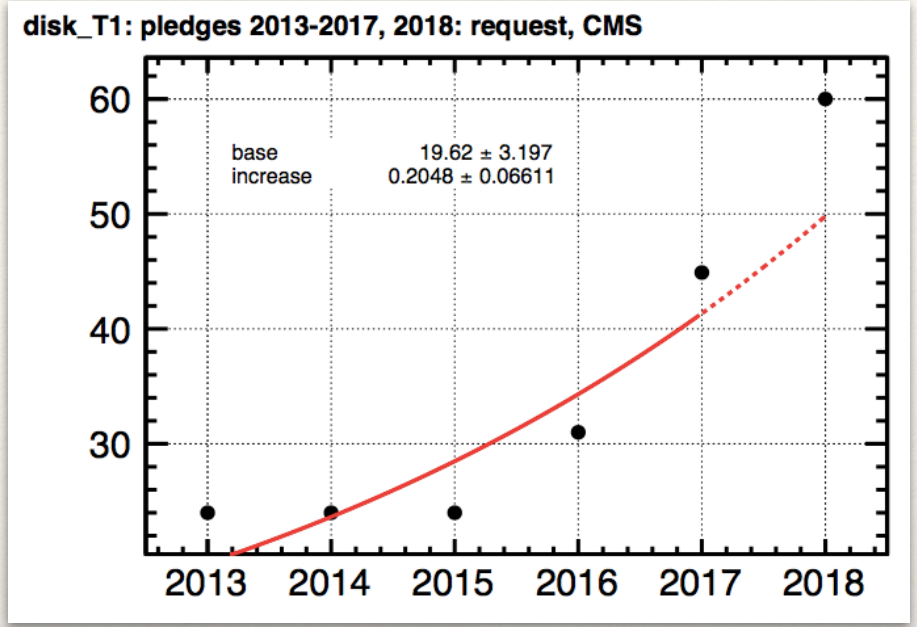


D. Lucchesi for C-RSG CERN-RRB-2017-057

April 25, 2017

Some examples from CRSG for CMS





Annual percentage increase in pledges

(from CRSG, fitting 2013-2017 WLCG REBUS data)

	ALICE	ATLAS	CMS	LHCb
CERN CPU	37%	44%	38%	23%
CERN disk	31%	29%	39%	35%
CERN tape	23%	39%	33%	43%
T1 CPU	25%	27%	39%	19%
T1 disk	27%	19%	20%	24%
T1 tape	32%	56%	36%	39%
T2 CPU	14%	29%	19%	38%
T2 disk	17%	14%	16%	24%
Averages:	~26%	~32%	~30%	~31%

flat budget?

General CRSG considerations

Resource requests for 2018

19

- During the October RRB FA made clear that 2018 could not be a(nother) special year.
- Experiments were asked by LHCC/WLCG to develop a strategy to mitigate resource requests without jeopardize physics.
- Final LHCC document reports: "*The LHCC notes that the margins to reduce the resource usage in the short term without impact on physics have been exhausted*"
- CRSG decided to evaluate 2018 requests respect to 2017 pledges

- CMS has a deficit of tape at T1
- CMS and ALICE have deficit of disk at T1 and T2
- CPU deficit is not crucial

	ALICE	ATLAS	CMS	LHCb
CERN CPU	0%	0%	0%	0%
CERN disk	0%	0%	0%	0%
CERN tape	0%	0%	0%	0%
T1 CPU	-8%	-12%	-14%	-4%
T1 disk	-14%	1%	-21%	-6%
T1 tape	-1%	-7%	-24%	-3%
T2 CPU	-24%	-13%	-7%	27%
T2 disk	-28%	-7%	-22%	-30%

CRSG scrutiny of 2018 resources

Resource	Site	2017 Pledge	2018 CMS	Growth	2018 CRSG	Growth
CPU (kHS06)	T0+CAF	423	423	0%	423	0%
	T1	515	600	17%	600	17%
	T2	791	900	14%	900	14%
Disk (PB)	T0+CAF	25	26	6%	26	6%
	T1	45	60	34%	60	34%
	T2	53	70	32%	70	32%
Tape (PB)	T0+CAF	71	80	13%	97	36%
	T1	133	205	54%	188	41%

w.r.t the 2017 pledges

(pledges, not requests here..)

CRSG acknowledged that CMS continues to have a deficit in disk and tape and acted as follows:

On **disk**, from the CRSG chair slides: "*Disk space requests for 2018 at T1 and T2 are above the general funding profile, we recommend them and since the CPU requests are below the flat budget expectations we ask the FA to help on disk.*"

On **tape**, from the CRSG chair slides: "*CMS is suffering from a deficit of tape space at T1s since few years. To mitigate the requests at T1s, that unlikely will be satisfied this year, we ask CERN to provide more tape space. Requests for CERN are CPU=0 and only +6% increase in disk space*".

CMS 2018 requests

As from the document submitted to CRSG:

w.r.t 2017 CRSG Oct'16

(approved requests here..)

Resource	Site	2017 CRSG (Oct16)	2018 CMS Request (Oct16)	2018 CMS Request (Apr17)	Rel. Increase (2018 Apr17 vs 2017 Oct16
CPU (kHS06)	T0+CAF	423	428	423	+0%
	T1	600	700	600	+0%
	T2	850	1000	900	+6%
Disk (PB)	T0+CAF	24.6	26.1	26.1	+6%
	T1	57	72	60	+5%
	T2	68	83	70	+3%
Tape (PB)	T0+CAF	70.5	91	80	+13%
	T1	175	230	205	+17%

Table 11: *CMS resource request for 2018. The first column shows CRSG Oct 16 recommendations for 2017. The second and third previous and current requests for 2018, with the relative increase between current request and 2017 CRSG recommendations shown in the last column.*

well within a flat budget

CRSG summary

CRSG congratulated "*WLCG and the experiments for the work done to mitigate the resource requests without jeopardising physics*"

CRSG encouraged all experiments "*to pursue use of non-WLCG CPU resources. To help monitor this, we recommend that all experiments quantify more fully the non-WLCG resources in their future reports*"

CRSG reports that "*efficiency and reduction of data stored on disk are almost at the limit. It is not clear that there is substantially more efficiency that can be gain without extensive reworking of the computing model*"

CRSG "*strongly support software engineering development and recommended that sufficient effort is funded to support this activity in the collaborations*"

CRSG conclusion on the flat-budget is that "*the assumption of a flat budget is not consistent with the historical pledge resources and we recommend a reevaluation of the assumptions of what a flat budget entails*"

Will 2018 be a problem?

If the FAs support us as close to the CRSG-endorsed numbers as possible, with our applied mitigations 2018 should be a relatively calm year:

- **Tier-0**: substantially unchanged on our side (CRSG suggested CERN helps us more with tapes - which is optimistically the easiest/cheaper resource category)
- **Tier-1**: CPU ok, thanks to premixing. Disk: driving down the # replicas is the key change. Tape: stop writing RECO (and remove), smaller GEN-SIM size and less GEN-SIM on tapes
- **Tier-2**: CPU needs would be stable, but reduced in the requests profiting from Tier-0 workflow optimizations. Disk: reduced needs due to fewer AOD copies

Looking forward

CMS towards the HL-LHC Computing TDR

Strategy document in 2017 as a roadmap towards the TDR in 2020

Primary input is the CWP process under the HSF umbrella

- which potential computing models (and related implications)
- which R&D needed, which demonstrators/prototypes
- build on the input from the cross-fertilising CWP working groups

CMS:

- involved and active in most/all the CWP working groups
- convergence with CMS-specific R&D areas in the Sw/Comp organisation
- input from CMS ECoM-17 (Evolution of the Computing Models) team

Resource tracking

As a community, we have a chance to improve on how we track the resources that may be available at any moment in the future

- why now? not a “plus” any more, now it starts to be a “must”
- (very informal and embryonal) first discussions among ATLAS and CMS

Items in random order (perhaps increasing complexity?):

- ability to get info on pledges insertions/modifications over time
- ability to track deployment updates/delays in various regions
- addition of info on possible beyond-pledges resources
- ability to specify time-limited resource availabilities (like 3 months of grant from X)
- cross-experiment connection among computing resources boards?
- tools to allow a coherent orchestration of diverse (owned vs not-owned) resources? is what we have the right set of tools to expand to opportunistic CPU cycles (as the CRSG has mandated)?
- (.. more ..)

CMS Analytics

A large and heterogeneous set of computing (meta)data sits in our databases

- despite growing and precious, it has rarely been accessed - until ~recently

Two examples of ad-hoc infrastructures (*no details today*)

- HTCondor classads into **Elasticsearch** (Nebraska, then CERN)
- main CMS computing data collections onto **HDFS**, run **Spark** jobs

One of the key activities to understand our performances

- brilliant sector to partition the work into student projects
- productive work with the WLCG UP team + excellent support from CERN-IT
- need to raise the awareness (e.g. need more disk for Elasticsearch at CERN)

Change in the CMS Sw/Comp mgmt

Daniele Bonacorsi's mandate ends on 31 August 2017

- **thanks to all LHCC colleagues for the fruitful collaboration in these exciting years!**

Please welcome Tommaso Boccali, in charge from 1 September 2017 for 2 years. He will co-coordinate the (now joined) Software/ Computing project with Liz Sexton-Kennedy.