Planning for CCRC´08/phase-2:

CMS

L. Bauerdick [Data Ops - CCRC co-coordinator]

D. Bonacorsi [Facilities / Infrastructure Ops - CCRC co-coordinator]

Outline

- Next months: the CMS schedule

- Tier-0 workflows
  - Cessy-CERN transfer tests
  - Tier-0 processing
  - CAF workflows
- Distributed data transfer tests
  - T0-T1, T1-T1, T1-T2, T2-T1
- Tier-1 workflows
  - Re-processing
  - Skimming
- Tier-2 workflows
  - MC production
  - Analysis

Major CMS exercises for 2008/Q12

- CCRC'08 (Feb 08, May 08)
  - WLCG Common-VO Computing Readiness Challenge (not only CMS)
    - Phase-1 for CMS: functionality and performance tests in Feb (4^{th} - 29^{th})
    - Phase-2: 4 week challenge in May (5^{th} - 30^{th})
      - CMS emphasis on a multi-VO stress test of the computing system/ops for data-taking at scale
- iCSA08 (May 08) (+CRUZET: see next slide)
  - "initial" CSA08
    - Focus on 1, 10 pb^{-1} scenarios
      - CMS offline workflows to check readiness for LHC data (calib, skims, express analyses)
  - Activities will run simultaneously with CCRC'08/phase-2
- CRAFT (Jun 08)
  - Cosmic Run at 4T
    - Extend the GRUMM + CROZET experiences (see next slide)
      - CMS emphasis on online+offline workflows to demonstrate readiness for data-taking ops
- fCSA08 (Jul 08)
  - "final" CSA08: further exercises for LHC collisions…
    - Focus on 10, 100 pb^{-1} scenario
  - …or the real thing!

Detector installation, commissioning and operation

Preparation of Software, Computing and Physics analysis

2008

Jan

Feb — CCRC'08-1

Mar — GRUMM

Apr — CRuZeT

May — iCSA08 / CCRC'08-2

Jun — pre- CR 4T

Jul — CR 4T (aka "CRAFT") — fCSA08

Aug

Sep

Oct

Cooldown of magnet

Private global runs (2 days/week) & Private mini-daq

Low i test

Beam-pipe baked-out

Pixels installed

CMS closed

CMS week in Cyprus

Initial CMS ready for run

2007 Physics Analyses results

CMSSW 1.8.0 sample production (*slipped into April*)

2 weeks of 2.0 testing

500 Mevts FastSim prod !

iCSA08 sample generation

CMSSW 2.0 release
[production start-up MC samples]

CMSSW 2.1 release
[all basic sw components ready for LHC, new T0 prod tools]

or beam!

Must keep exercises mostly non-overlapped

iCSA is part of CCRC/phase-2

iCSA goals are those of a CMS-specific exercise strongly tight to the commissioning/physics schedule

- Preparation for iCSA08 is proceeding and has priority wrt CCRC during May…
- … but "data" has priority over iCSA !
  - completely driven by requirements of CMS Commissioning and Physics
    - "play through" first 3 months of data taking
    - commissioning/physics requirements become clearer week after week…
  - iCSA will put stress on 'some' aspects of our computing systems
    - simulation production ≤May, across many Tiers
    - To and CAF at CERN: prompt reconstruction, AlCa processing, CAF-based exercises
    - T1's: reprocessing
    - T2's: no big plans in iCSA to cover all T2's for real analysis running, but "some" plans

CCRC goals are foremost that of a multi-VO computing scale test

- CCRC activities to augment iCSA loads with additional computing tests
- ensure all systems are being stress tested end-to-end to prove metric
- should include planned tests of Tier-0 components, e.g. repacker tests, etc.
- important to coordinate carefully with iCSA, which has priority

CCRC/phase-2 areas of work

To prove that systems are ready for **continuous** and **simultaneous** use via **complete workflows** by **CMS and other LHC exps** at **"full" scale**

This requires, in addition to iCSA, to fully include tests of:

- To workflows
  - Cessy link to To
  - tests of new Tier-0 components when they become available
  - CAF workflows
- Distributed data transfer tests
  - watch for interference w/ other experiments, especially at the To and T1 level
    - Transfer systems, tape drives, … shared across experiments
- T1 skimming operations during handling of incoming data, re-processing
  - stringent test of storage system performance of simultaneous reads and writes
- T2 analysis to produce "chaotic" workload across the systems
- Monitoring, to establish metrics

Other important computing goals for CCRC

- Include all sites which have to be ready for the '08 run
- Kind of "test" of the affiliation of Tier-2 sites with Analysis Group
- Achieve analysis metric and prove that the system can sustain load from analysis users

Use existing computing teams, run in "operations" mode

- Data Ops, Facilities Ops, PADA team

In the following:

- *Go through CMS/CCRC areas of work in May, one by one*
- *Discuss what iCSA does and what CCRC does*

- *Give info on proposed CCRC approach, goals and metrics*

  Whenever available. Details not in this talk will come within this week

Simulation data for analysis

- Production of simulation samples with CMSSW 2.0

iCSA08 | CCRC'08-2

Focus on 2 scenarios for 2008 data-taking:

"**S43**": 43×43 bunches, L~2×10^{30}, 1 pb^{-1}, O(150M) events

"**S156**": 156×156 bunches, L~2×10^{31}, 10 pb^{-1}, O(150M) events

expect production of <150 M events for May but S43 and S156 will have a lot of overlap

Conditions: No pile-up; assume a complete detector; zero suppression

Generator streams

physics samples (minbias, jets, leptons) + technical samples (cosmics, …)

Assume Storage Manager bandwidth can sustain 300 MB/s, and allocate:

~100 MB/s for specialized calibration stream, ~200 MB/s for normal sized events

Trigger menus to be defined for L=2×10^{30} and 2×10^{31}

Define a simple set of Primary Datasets (PD) from generator streams

PD's defined by HLT; no mixing of generator streams to make PD's

Same for technical streams + …

<u>Where are we now?</u>

First: planning and discussing with commissioning/physics stakeholders

Now: pre-production started!

To workflows:

Cessy-CERN transfer tests

To processing

CAF workflows

Mike Miller, Dave Mason, Peter Kreuzer, Stephen Gowdy

Cessy→CERN transfer tests

- P5→T0 link commissioning

iCSA08 | CCRC'08-2

Receive events in RAW format from Storage Manager

Important to test the 10 Gb/s link Cessy-CERN separately from other T0 tests

Goal: load up to 300 MB/s of sustained throughput

Current plan calls for a CMS readout test, targeting at 300MB/s, for the whole period (aka: until it breaks)

Commissioning readout of the full detector will run concurrently to P5 link commissioning

CCRC will recycle a simulated data sample to augment the existing traffic as needed, with a lower priority

Where are we now?

First successful transfers already during CCRC'08/phase-1 tests in Feb

Plan to demonstrate target rate and sustainability

Processing at CERN

- Prompt reconstruction — iCSA08 | CCRC'08-2
  1. Prompt reco (pass 1) of S43 (150M evts) with start-up AlCa conditions
     - As soon as sizeable data sample is generated
  2. Prompt reco (pass 1) of S156 (150M evts) with improved conditions based on first 1 pb^{-1}
     - Production schedule allows next batch of 150M events in fCSA08 only
     - Constants prepared beforehand, or take from exercises if possible
       - (e.g. ECAL+HCAL calibration, Tracker and muon alignment)

  Estimated duration: 2 weeks
- Tape writing exercises — iCSA08 | CCRC'08-2
- Planned tests of new To components — iCSA08 | CCRC'08-2
  - Merging and repacking tests at the To, two-file output (RAW+RECO), …

Where are we now?

Reconstruction at To and tape writing successful in CCRC'08-1
- Need for a full multi-VO approach, still

Some technical aspects
- Test of application of large conditions payloads at To (and T1)
- Be sure the Castor set-up at CERN is the appropriate one (To vs CAF)

Need development still in new components
- Test what's available in May

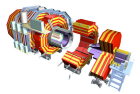
Processing at CERN, and CAF

- Low turn-around calibration & alignment exercises — iCSA08 CCRC'08-2
  - Aim to demonstrate as much of the complete workflow in as realistic a manner as possible
    - Development/test of any needed L1/ HLT triggers to select evts for calib stream
    - Realistic AlCaReco skims run at To to produce reduced datasets for AlCa jobs
    - Exercises to produce new AlCa focusing on early LHC data samples
  - In general extending the "proof-of-principle" demonstrations from CSA07 to be ready for LHC data
- Coordinated, quick-response physics analyses — iCSA08 CCRC'08-2
  - For the CMS start-up, it is having prompt access to all types of data (muons, electrons, jets) that is most important to validation efforts
    - not full statistics and not just one sample
    - Propose that we direct at least some % of each dataset to CAF (maybe 100% of them…)
- Ramp-up resources (see next) — iCSA08 CCRC'08-2
- Ramp-up the scale of users (see next)

Where are we now?

- First interesting CAF exercises just at the end of CCRC'08-1
  - More insight about the needed work on the CAF came from GRUMM
- Need to review the overall access pattern, the policies, and the implementations

Progresses on CAF resources and set-up



LSF CPUs : currently 328 cores (2008 Pledge : 1200 cores)

328 job-slots available +1000 slots allocated to To this week, available to CAF on demand

Group allocation : ALCA : 40% , PHYS : 20% , COMM : 40%

Maybe need to isolate time-critical ALCA queues from rest ?

CASTOR "cmscaf" pool : 263 TB (2008 Pledge : 1.6 PB)

upgrade to 500TB by end-April, further upgrade sometime (?) during May

write permission restricted to CAF LSF-users + cmsprod + phedex

planed group allocation : ALCA: 10%, PHYS: 40%, COMM: 50%

would require Service Classes, under discussion

Interactive CAF servers: 40 cores

2008 pledge : 400 cores, but very modest utilization so far

Interactive /afs user space: 3x150 GB /afs space ready

/afs/cern.ch/cms/CAF/CMSALCA

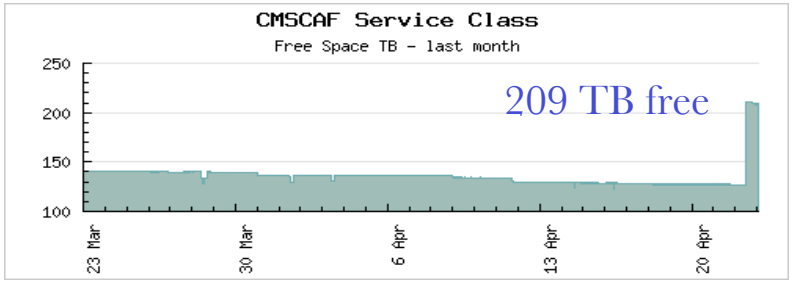
/afs/cern.ch/cms/CAF/CMSCOMM

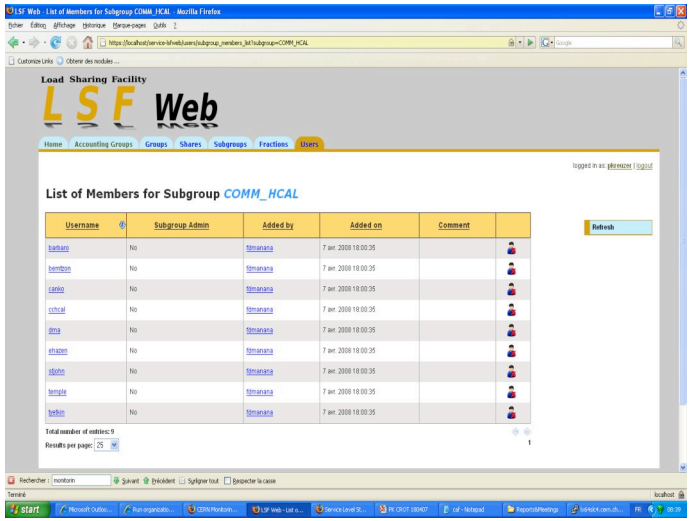
/afs/cern.ch/cms/CAF/CMSPHYS

Created CAF sub-group directories

ACL's, user quotas to be managed by (sub-)group leaders

Automated interface to LSFWEB user-list in place





iCSA resources table

	CPU	Disk	Tape
CAF 2007	128 slots	35 TB	
CAF Feb-Mar08	228 slots	172TB	
CAF mid-Apr08	328 slots	263 TB	
CAF iCSA08	328 + (1000)slots	500 TB	
CAF 2008	1200 slots	1600TB	3 PB
To mid-Apr08	2000 slots	350 TB	
To 2008	3000 slots	400TB	

+ dedicated servers on CAF : 2 Millepede Servers , 1 CRAB Server

Plans for iCSA and beyond

Automated PhEDEx Transfer Subscriptions To-CAF

- Plan to have cron job running on To to make automated subscriptions
- details to be worked out among PhEDEx experts (timescale: 1 week)

Automated CAF job triggering

- For iCSA08 : minimum is web summary page with incoming data
- PhEDEx agent: checking last file in a block and trigger CAF jobs
- Long term: WMBS/CRABServer solution

CAF workflow integration in CRAB Server

- Good progress on integrating first ALCA workflow into CRAB Server
  - Based on 1_8_4 FastSim AlcaReco

Systematic Monitoring of cmscaf queue and castor pool

- To optimize resource utilization and monitor stability
- To be soon displayed live in CMS Offline center

Distributed Data Transfers

Douglas Teodoro, Paul Rossman

Distributed Data Transfers

PhEDEx-driven FTS transfers:

- T0→T1, T1↔T1, T1→T2 and T2→T1

iCSA08 CCRC'08-2

Where are we now?

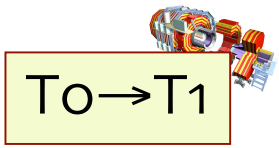
Highly exercised in CCRC'08-1

Continuously exercised before, during and after CCRC'08-1 via the DDT program

Need to:

- ramp-up the scale (up to 2008 full rates)
- check the multi-VO factor
  - e.g. To-T1: watch for interference w/ other exps (tape systems, tape drives shared across exps)
- check the multi-activity environment
  - e.g. reproc + transfers at T1's: stringent test of storage system performance of simult reads and writes
- extend
  - e.g. define a policy and run non-regional T1-T2 transfers

Distributed Data Transfers



3 categories of data will be flowing from CERN to T1's

- iCSA "custodial"

  iCSA08 | CCRC'08-2

  - DataOps is authoritative in defining those, FacilitiesOps checked these are not behind tape space declared available on May 1st, CMS people at T1 site can complain if disagree, these data must be kept beyond end of May
- iCSA "not-custodial"

  iCSA08 | CCRC'08-2

  - CMS people at T1 sites can decide what to get in addition to custodial, these data can be deleted on Jun 1st. T1's are not asked to send these to tapes.
- CCRC "LoadTest data"

  iCSA08 | CCRC'08-2

  - Designed and managed to augment/customize the traffic load as needed, via usual LoadTest infrastructure and an extended LoadTest suite to be used for pure computing scale tests, can be deleted once on tapes, even within May. Shares and LT mgmt are dealt with by ops team.

Different treatment:

- "custodial"
  - Get from DataOps a list of LFN's, follow the namespace recipe and define tape families accordingly, then CMS contacts at T1 sites test them before first data arrives
- "non-custodial" + "LoadTest data"
  - Not asked to send them to tapes: just do map them both onto a unique tape family

Proposal of running a deletion exercise on "non-custodial" data for 10-15 days in first half of June

iCSA08 | CCRC'08-2

- Under discussion

Distributed Data Transfers

iCSA08 | CCRC'08-2

Goal: exercise the To data export to the T1's

Tools: use the LT suite

Preparation:

- Central ops team extend the LT sample, inject, manage,…
- All these data go to tapes, so T1's will define a tape family for LFN's like "/store/CCRC08LoadTest/.* "

Testing time window: check with ATLAS for the best superimposition. CMS would like to have these tests running during the whole CCRC/phase-2

Proposed metric:

Participation

Aim is to have all 7 T1's in the game for most of the May 5-30 time window

Throughput (2008 target based)

Go back to megatable, re-arrange back the ASGC load redistrib, and get 2008 figures

Demonstrate rates for 3 days in a row at least

100% (**extra green**), 75% (green), 50% (acceptable), <50% (failed)

Hence, a clear progress wrt Feb should be demonstrated *(i.e. doing 40% is not enough any more)*

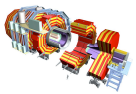
Continuous import (100% = all "usable" days in the May 5-30 window)

Evaluate T1 ingestion days *normalizing* to acceptable conditions for ops

100% (**extra green**), 75% (green), <75% (failed)

Measurement of the *transfer latency* is a target (not only for the To-T1 route, btw)

We will not put threshold based on this, but we plan to get a clear measurement of it out of CCRC



S43 + S156	Events	RAW Size/ event [MB]	RAW size per request [TB]	RECO size/ event [MB]	RECO size per request [TB]	Total size per requestGen + 4 RECO [TB]	Custodial Site Allocation
Minbias	25,000,000	0.63	15.75	0.32	8.00	47.75	FZK
Jet20	4,000,000	1.00	4.00	0.50	2.00	12.00	CNAF
Jet30	4,000,000	1.00	4.00	0.50	2.00	12.00	ASCG
Jet50	4,000,000	1.50	6.00	1.00	4.00	22.00	FNAL
Jet80	4,000,000	1.50	6.00	1.00	4.00	22.00	FNAL
Jet110	4,000,000	1.50	6.00	1.00	4.00	22.00	PIC
Muon5	10,000,000	1.00	10.00	0.50	5.00	30.00	CC-IN2P3
MuonCosmicBON	10,000,000	0.25	2.50	0.10	1.00	6.50	RAL
MuonCosmicBOFF	10,000,000	0.25	2.50	0.10	1.00	6.50	RAL
TrackerCosmicBON	10,000,000	0.25	2.50	0.10	1.00	6.50	CNAF
TrackerCosmicBOFF	10,000,000	0.25	2.50	0.10	1.00	6.50	ASGC
HaloMuon	10,000,000	0.25	2.50	0.10	1.00	6.50	PIC
TrackerHaloMuon	10,000,000	0.25	2.50	0.10	1.00	6.50	PIC
HCALNZS	10,000,000	2.50	25.00	1.50	15.00	85.00	FNAL
HCALISOTRACK S43	2,000,000	1.50	3.00	1.00	2.00	11.00	CC-IN2P2
	127,000,000		**94.75**		**52.00**	**302.75**	

S156	Events	RAW Size/ event [MB]	RAW size per request [TB]	RECO size/ event [MB]	RECO size per request [TB]	Total size per requestGen + 2 RECO [TB]	
Jet150	4,000,000	1.50	6.00	1.00	4.00	14.00	CNAF
Muon11	10,000,000	1.00	10.00	0.50	5.00	20.00	RAL
MuonOnia	1,000,000	1.00	1.00	0.50	0.50	2.00	ASGC
LeptonEWK	200,000	1.50	0.30	1.00	0.20	0.70	ASGC
HCALISOTRACK S156	2,000,000	1.50	3.00	1.00	2.00	7.00	ASGC
	17,200,000		**20.30**		**11.70**	**43.70**	

	S43+S156 [TB]	S156 [TB]	Total [TB]
T0 prompt RECO 0pb calibration [TB]	52.00		52.00
T0 prompt RECO S43 calibration [TB]	52.00	11.70	63.70
ReReco S43 calibration [TB]	52.00		52.00
ReReco S156 calibration [TB]	52.00	11.70	63.70
	302.75	**43.70**	**346.45**

Site	WLCG 2008 pledges [TB]	WLCG 2008 pledges [%]	Renormalized Pledges [%]	iCSA08 custodial [TB]	Crosscheck, sum of allocated samples [TB]
ASGC	585	6	10	34.65	28.20
CC-IN2P3	866	9	12	41.57	41.00
CNAF	735	8	10	34.65	32.50
FNAL	4700	51	35	121.26	129.00
FZK	900	10	13	45.04	47.75
PIC	731	8	10	34.65	35.00
RAL	668	7	10	34.65	33.00
	9185	**100**	**100**	**346.45**	**346.45**

Comparing with the tape space available today, or the pledges as of May 1st, CMS is confident that T1's will not have problem in ingesting the custodial shares

All additional, non-custodial traffic will come from scale test needs

Reviewing the gigatable input nbs → more info within this week

Distributed Data Transfers

iCSA08 | CCRC'08-2

Goal: verify the latency for AOD replication

Tools: use the LT suite

Preparation:

- Need to determine based on the megatable the size of the datasets for each T1
- Inject fake (LT) AOD datasets on each T1 based on the nominal AOD fraction
- Subscribe these to the other 6 T1's
- Run agents (not altogether but by rotation) and monitor the replication (sharing the load via re-routing)
- The target will be replication within 2 weeks.•.•The LFN should be e.g. '/store/PhEDEx_LoadTest07/.*CCRC08_phase2_AOD_T1_X.*'."
- No tapes

Testing time window: check with ATLAS for superimposition. CMS would like to have these tests running the whole month - maybe adding 1 T1 at a time in the first 2 weeks, and let them run each for 2 weeks

Proposed metric:

Participation

Aim is to have all 7 T1's running their own AOD replication exercise

Latency

Target is to demonstrate it's possible in approx 2-weeks

AOD syncronization pattern

Analyze how the transfer pattern is (continuity, quality, throughput, …)

Distributed Data Transfers

T1 → regional T2

proposed CCRC metric

iCSA08 | CCRC'08-2

Goal: exercise the T1 data export to regional T2

Tools: use the LT suite

Preparation:

- Nothing different wrt usual LT/DDT, but rate

Testing time window: each region decides on 1 week + another (repetition) week (if needed); week 4 stands as default repetition week for all

Proposed metric:

Participation

Only the T2's in the region with an "OK" from PADA Site Commissioning and DDT (e.g. DDT link commissioning ok, SAM tests ok) are allowed to try, and aggregated targets will be evaluated

Throughput (2008 target based): aggregate targets

for 3 days in a row

100% (**extra green**), 75% (green), 50% (acceptable), <50% (failed)

Continuous T1 downstream traffic (100% = all "usable" days in the testing week)

Evaluate T2 ingestion days *normalizing* to acceptable conditions for ops

100% (**extra green**), 75% (green), <75% (failed)

Distributed Data Transfers

T1 → non-regional T2

proposed CCRC metric

iCSA08 | CCRC'08-2

Goal: exercise the T1 data export to non-regional T2

Tools: use the LT suite

Preparation:

- PADA/Integration agreed that this step may be handled directly by the DDT task force
- DDT-TF may run ad-hoc rotation scheme + slightly CCRC-adapted monitoring tools

Testing time window: all 4 weeks

Proposed metric:

Participation

Only the T2's in the region with an "OK" from PADA Site Commissioning and DDT (e.g. DDT link commissioning ok, SAM tests ok) are allowed to try

Throughput (2008 target based):

for 2 days in a row

>75% (**extra green**), 50% (green), 40% (acceptable), <40% (failed)

Distributed Data Transfers

Goal: exercise the T2 upload to associated T1

Tools: use the LT suite

Preparation:

- Nothing different wrt usual LT/DDT, but rate

Testing time window: each region decides on 1 week + another (repetition) week (if needed); week 4 stands as default repetition week for all

Proposed metric:

- Participation
  - Only the T2's in the region with an "OK" from PADA Site Commissioning and DDT (e.g. DDT link commissioning ok, SAM tests ok) are allowed to try
- Throughput (2008 target based): aggregate targets
  - for 3 days in a row
  - 100% (**extra green**), 75% (green), 50% (acceptable), <50% (failed)

T1 workflows:

Re-processing
Skimming

Oliver Gutsche, Guillelmo Gomez Ceballos, N.N.

Re-reconstruction at T1's

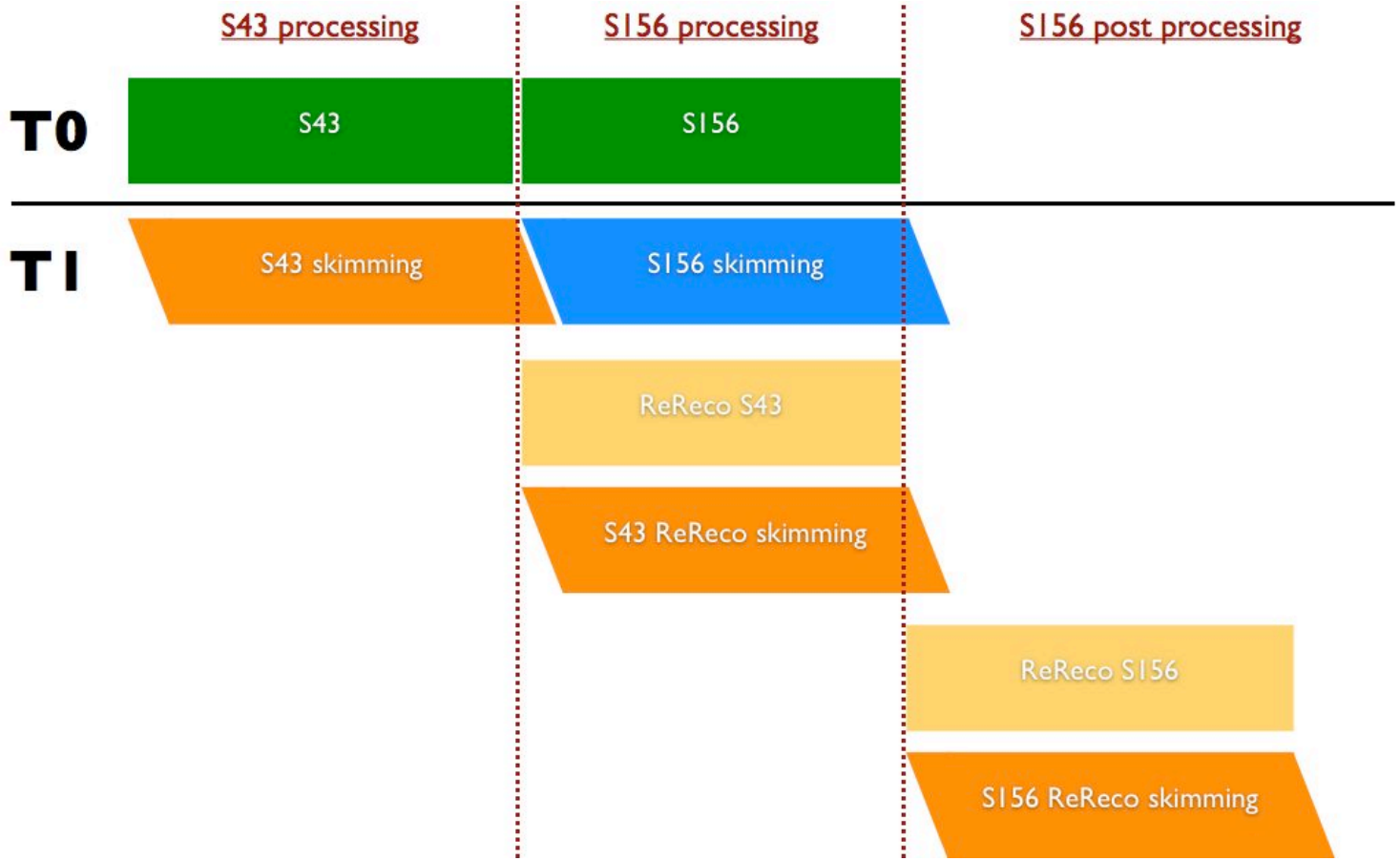
iCSA08 CCRC'08-2

- Apply new calib constants as data arrive at T1's from To
  - S43 (pass 2): with improved constants based on 1 pb^{-1}
    - i.e. re-reco of S43 with S43-based AlCa
  - S156 (pass 2): with improved constants based on 10 pb^{-1}
    - i.e. re-reco of S43 and S156 with S156-based AlCa
  - Estimated duration: 2 weeks
  - Current proposal
    - Monitor reco at T1; analyze at T2's
    - We may need to send the RECO back to the CAF after re-processing at T1's
- Check the scale and the multi-VO factor

iCSA08 CCRC'08-2

Where are we now?

- Re-reco at T1's was closely tested in recall/reproc exercises in CCRC'08-1
  - Many crucial technical issues here, still to be fully addressed, e.g. tape families
  - Also: push more for a multi-VO approach, and measure



Processing metrics *Work in progress…*

Metric	Description	Target	Monitoring source
job submission	number of jobs that can be submitted and in the queue within on hour	5500	ProdMon
job throughput	number of jobs per day	22k	ProdMon
job saturation	sustain usage of all T1 slots for a period of time	7 days	ProdMon
processing latency	latency to finish processing block, migrate to global DBS and insert into TMDB (takes into account processing problems, averaged over all processed blocks, subtract processing time)	?	ProdMon & TMDB
DBS migration latency	time to migrate processed block to global DBS	?	?
DBS query latency	time to query DBS (needed for continuous processing when blocks arrive at sites)	?	?

T2 workflows:
MC production
Analysis

Frank Wuerthwein, Alessandra Fanfani, Ken Bloom, Giuseppe Bagliesi

Production and Analysis at T2's

Production, in preparation for iCSA08

iCSA08 CCRC'08-2

- Use all Tiers for GEN-SIM-DIGI-RAW-HLT production
  - Also T1's may be of help
- Pre-production running NOW.
  - Basically: same as normal MC prod, but more intense
  - DataOps estimate: ~8k slots (T1+T2) used permanently
    - T2's will be at least running FastSim MC for MadGraph production at 10 TeV

Analysis

- Possible activities in a iCSA context are: iCSA08 CCRC'08-2
  - Repeat at T2's the CAF-based physics analyses with re-reco data
  - Repeat at T2's the CAF-based AlCa exercises
- More activities needed in a CCRC context iCSA08 CCRC'08-2
  - T2 analysis to produce realistic "chaotic" workload across the systems
  - Requires additional help from CMS people at T2 sites
    - Propose to set-up standardized jobs, to be run by CMS physicists close to T2's
    - Help from T2 data managers (and beyond) to transfer data and run jobs
    - Need to stress-test T1-T2 transfers + job submission infrastructure + site ops

Analysis exercises at scale at T2's

Goal: gain an overall understanding of the performance and readiness of the T2's globally for CMS data analysis

Phase-0: preparation(s) phase

Phase-1: "controlled" job submissions

Phase-2: "chaotic" job submissions

Phase-3: "stop-watch" phase

An approach slightly more focussed to physics preparation indeed may be plugged into Phase-1

See next slides - Step-A/B

Phase-0: preparation(s) phase

Prior to May, to get ready

- Datasets for sites to pull
  - Similar in physics content, different in size
  - Choose the appropriate size for the size of your T2
  - Pull it via PhEDEx down to your T2
- Executable will be on CVS, build and CRAB instructions also
  - you (T2) play with it before CCRC
- Policy config will be given
  - User area, DN which should write there, …
- Declarative twiki area for each T2
  - How much disk may we use there?
  - How many slots do I expect to be able to access there?

Phase-1: controlled job submissions

Single submission point of standardized CRAB jobs

- Prepared in advance ("crab --create …")
- On all datasets that a given T2 has pulled
- Record success rate, I/O perf, …

Different jobs, increasing complexity, may not run all of them at all T2's

1. Long-running CPU-intensive job with moderate IO requests
2. Long-running IO-intensive job
3. Short (~10 mins) IO-intensive job

Use srm-ls from remote to check the O(10 MB) output

Report the observations in running these tests on T2's.

Phase-2: "chaotic" job submissions

Encourage T2 people to run the 3 jobs themselves on T2's

Configure the stage-out to your 'home' T2

Datasets distributed, users have their disk space at a T2

Monitor with the Dashboard

Phase-3: "stop-watch" phase

Measure the total latency of T2 ops

- Identify a set of datasets located across all T1's
- ask T2's to pick some, subscribe to them, download them, run jobs on them, measure performance and success vs failure rates, and the total latency from the beginning to the end

Analysis exercises at scale at T2's

An added value may come if the exercises closely mimic (at scale) the use of analysis tools by 'sample' Physics Groups

- Define 3 fake physics groups: fakeQCD, fakeEWK, fakeHiggs
- Associate a list of T2's to each fake Physics Group (PG) based on declared affiliations, whenever possible
- Use CRABSERVER to submit fake-but-realistic PG tasks

A test plan is in place for this, as follows.

Step-A: job type definition + run/measure

Which type of job?

- CPU-intensive job reading RECO and writing a user-defined rootuple
  - Start from existing analysis code, e.g. the QCD Underlying Event CMSSW16x analysis run on Njet samples produces a rootuple of about 3.5 KB/evt
- The job time lenght can be tuned adding more computation in the job
- Remote stageout of rootuple to 1 T2 or small subset of T2's associated to PG
  - A periodic (srm-rm) cleanup of the produced rootuples to avoid pollution

For each fake Physics Group:

- Submit jobs to affiliated T2's
  - Rough (and upper limit) estimate based on declared T2's pledges for analysis (*) for a job of 3 hours are show on the table on the right
    - (*) https://twiki.cern.ch/twiki/bin/view/CMS/CMST2Pledges
  - Measure % of failures at each T2 affiliated to that fake PG
  - Decouple the % of failures due to remote stageout
  - Measure the latency to get back the results for that T2
    - distribution of time spent from job submission to end of job
- (… continues…)

```
Total of 44085.0 jobs in 24 hours
over >20 sites

T2_UK_SouthGrid jobs = 750.0
T2_BR_UERJ      jobs = 2145.0
T2_TW_Taiwan    jobs = 600.0
T2_CN_Beijing   jobs = 135.0
T2_DE_DESY      jobs = 1800.0
T2_CH_CSCS      jobs = 720.0
T2_DE_RWTH      jobs = 1350.0
T2_UK_London    jobs = 2150.0
T2_IT_Bari      jobs = 1080.0
T2_BE_UCL       jobs = 850.0
T2_ES_IFCA      jobs = 1800.0
T2_US_All       jobs = 18750.0
T2_FR_CCIN2P3   jobs = 875.0
T2_IT_Legnaro   jobs = 1000.0
T2_ES_CIEMAT    jobs = 835.0
T2_KR_KNU       jobs = 2000.0
T2_HU_Budapest  jobs = 645.0
T2_FR_GRIF      jobs = 2000.0
T2_BE_IIHE      jobs = 500.0
T2_IT_Rome      jobs = 700.0
T2_IT_Pisa      jobs = 2500.0
T2_BR_SPRACE    jobs = 900.0
```

Step-B: test the local-scope DBS scenario

- Define 1 analysis made of running on N samples at M sites
  - N to mimic running on bkgs+signal
  - M can be the whole set of T2's associated to the PG or a subset
- Group the result for each analysis with:
  - Measure % of failures with breakdown at each site
  - Measure the latency to get back the results

Assume 2 iterations per week per Physics Group

Then, pick just one of the fake PG's:

- on all(most) T2's affiliated, run jobs that write files with remote stageout only to 1 or 2 T2's of the group
  - register the files in private DBS
  - run analysis jobs that read files in private DBS

Possible timeline

- Assuming that later in the month real user CSA08 activities will take place, it may be wise to plan that before that happens
  - May week 1 or 2: *step-A*, with 2 iterations per week
  - May week 3 : *step-B*

Monitoring
and establishing the metrics

Stefano Belforte, Lassi Tuura

Monitoring the progress

Focus is on:

- Monitor activities in each testing block
- Establishing the end-to-end metrics

Work in progress on:

- Refresh of 'a-la CSA06' daily reports (useful also outside CMS)
- Aggregate plots of **/Prod** and **/Debug** instances
- Bi-dimensional rate/volume + quality plot, with metrics
- Latency plots (for distributed transfers)
- Buffer-2-MSS plots (from PhEDEx FilePump logs)
- # datasets/datablocks per T2, runtime plot
- T1 AOD load-sharing replication model

Automatic procedures whenever possible

Coarse approach is just fine, e.g. refresh frequency: ≥24 hrs

- Still OK for reporting overall progress. Refinements can come during/after May

Site people will receive CMS reports during CCRC

- to help them to understand what happens from the CMS pov

… but on CMS activities running at the site they are encouraged to use the standard production-quality monitoring tools, as usual, as before.

Summary

We are preparing to run major CMS exercises in May-Jul 2008

- May 08 is our next 'hot' period

Constructive superimposition of iCSA and CCRC/phase-2 is crucial

- iCSA as a CMS-specific test with a commiss./physics delivery attached
- CCRC as a multi-VO computing scale test

Sites (and site people) are our most precious resource

- T1 sites are establishing the CMS workflows
- T2 sites are becoming more and more crucial for CMS analysis

Infrastructure to run CCRC/phase-2 is mostly in place

- The missing pieces in tools will be finalized within end-April
- The missing pieces in planning will be finalized this week and early next one
- Support channels are being refreshed

Communication flow must be designed with care

- Among CMS computing sub-projects, and with WLCG
- All Tiers must be equally reached, e.g. work to improve coverage of A-P region